PARAMETERS OF THE ANALYTIC VECTOR REPRESENTATION OF SPEECH

A Thesis Presented for the Degree of

Doctor of Philosophy

in Electrical Engineering

in the

University of Canterbury

Christchurch, New Zealand

By

A.C.J. Hamilton, B.E. (Hons)

1985

ABSTRACT

The instantaneous amplitude (magnitude) and instantaneous frequency (phase derivative) waveforms of analytic vector representations of speech, models of speech and sub-bands of speech are generated and analysed.

Important characteristics of the instantaneous parameter waveforms are identified. These are related to attributes of vector loci and the spectral and zero structure of the associated real and analytic signals.

Real time generation and display of the analytic vector and its parameters is achieved and associated difficulties identified.

The problem of reconstruction of speech from its instantaneous parameters and from modified instantaneous parameters is addressed. Resulting distortions are classified and an application to low bit rate speech transmission investigated.

ACKNOWLEDGEMENTS

I would like to acknowledge the financial and technical support of the New Zealand Post Office provided during the course of this research and thesis preparation.

I also thank my supervisor, Mr J.A. Webb of the Electrical and Electronic Engineering Department, University of Canterbury, for his guidance and enthusiasm in the pursuit of this work. Many thanks must also go to my fellow postgraduate students, especially B.G. Henderson.

Finally, I acknowledge the major contribution of my wife, Julie, who spent many hours typing this thesis, and thank her and the rest of my family for their encouragement and support throughout.

TABLE OF CONTENTS

e. Rectangular Bandwidth Construction /ʃ/

f. Improved Construction /ʃ/

PART 2 TEST PHRASE (1)

a. Original

b. a(t) lowpass to 500Hz, full bandwidth $\omega_i(t)$

c. Full bandwidth a(t), $\omega_i(t)$ lowpass to 500Hz

d. Both a(t) and $\omega_i(t)$ lowpass to 1000Hz

e. Both a(t) and $\omega_i(t)$ lowpass to 500Hz

PART 3 APERIODIC VOWEL

a. Original

b. Both a(t) and $\omega_i(t)$ lowpass to 1000Hz

PART 4 TEST PHRASE (2)

a - n As detailed in Table 5.1

PREFACE

Speech transmission techniques which make use of "redundancies" inherent in the speech signal are normally classified into two basic types. These are Vocoders and Waveform coders.

Vocoders make use of a speech production model and transmit only information which is assumed to be required by the ear for perception. It is usually admitted that vocoded speech quality is "synthetic", even when high bit rates are used in transmission.

Waveform coders generally rely on redundancies provided by the statistics of a digital representation of the speech waveform. These coders provide high quality speech transmission, but at high bit rates and any bit rate reduction is at the cost of reduced speech quality or increased coder complexity.

Although normally classified as vocoders, I define a third type of speech coder which emcompasses all those which employ "frequency division". This group includes the Analytic Signal Rooter and the Phase Vocoder. None of this type of speech coder employ a vocal tract/excitation voice production model but do transmit only that information which is deemed necessary for perception.

The point of similarity between all frequency division vocoders is that sub-bands of speech are processed to obtain the parameters of their analytic vector representations. These instantaneous amplitude and instantaneous frequency waveforms are then further processed, transmitted and recombined on reception to form a version of the original speech. Although a surprisingly wide range of speech qualities are claimed

for output at various bandwidth reduction factors, particular types of speech distortion appear common to many frequency division vocoders. It is suggested that this is due to the techniques used to process and code instantaneous frequency.

The purpose of this research project was to generate and investigate the time waveforms which represent the magnitude (instantaneous amplitude) and phase derivative (instantaneous frequency) of an analytic vector representation of speech. It was hoped that these waveforms could be interpreted as meaningful parameters of speech and, as such, that they may exhibit forms of redundancy which would make them suitable for use in a bandwidth efficient speech transmission system. The following thesis documents the resulting research and conclusions.

The remainder of this preface is a chapter by chapter review of the thesis. All of the experimental hardware and software described was designed by myself and hardware constructed by myself or by technicians under my supervision.

Chapter 1 is the result of a literature survey of speech coding techniques. The speech signal is studied and the characteristics of basic speech elements (phonemes) listed. Perception of speech is investigated in terms of the performance of the ear and the importance of various speech parameters to intelligibility and recognisability. Bandwidth efficient speech coding techniques are studied in overview, with concentration on some typical systems and several types of frequency division vocoder. A scale of synthesised speech quality is introduced and used to rate the performance of several speech coding schemes.

Chapter 2 introduces the analytic signal, defines its properties and infers some properties of the instantaneous parameters. The instantaneous parameters of models of speech

sub-bands are then investigated. A sub-band of a voiced phoneme is constructed from a pair of sinusoids and resulting instantaneous parameters examined and related to vector loci. A sub-band of an unvoiced phoneme is approximated by narrow band Gaussian noise and the probability density functions (pdfs) of its instantaneous parameters investigated. Schemes suitable for resolving a baseband signal into its instantaneous parameters are then examined and this leads to representation of the instantaneous parameters and real signal in terms of the zeros of the associated analytic signal. By means of an example, I then show the relationships between analytic signal zeros, instantaneous parameter fluctuations, features of the analytic vector loci and the zero crossings and complex zeros of the real signal.

Chapter 3 covers the design, construction and operation of hardware to perform real time complex demodulation of baseband speech according to the equations

$$a(t) = (s^2(t)+\hat{s}^2(t))^{\frac{1}{2}}$$

and

$$\omega_i(t) = {}^{d}/_{dt}\{\tan^{-1}(\hat{s}(t)/s(t))\}$$

Features of the system include a real time Hilbert transformer based on FIR filter techniques and a "digital pipeline" computer to perform the complex demodulation. The design of each section of the system is fully documented. Displays of system output are by means of oscilloscope photographs and chart recordings and include vector loci, instantaneous parameter waveforms and time averaged instantaneous parameter waveforms. Output is illustrated for test waveform input and for various phonemes and a phrase.

Chapter 4 documents the analysis of results obtained using a computer based version of the complex demodulator designed in Chapter 3. Increased sampling rate and digital word length leads to high fidelity computer generated output. Analyses of full bandwidth and sub-bands of vowels are

interpreted to relate the instantaneous parameters to features of their Fourier amplitude spectra. This process is repeated for bandlimited unvoiced fricatives and attempts are made to classify unvoiced fricatives solely in terms of parameters of the pdfs of their instantaneous frequency functions. Analysis of a voiced fricative is based on results obtained for voiced and unvoiced fricatives and features of the instantaneous parameters are shown to be hybrids of these previous two types. Finally, the computer based complex demodulator is used to analyse three words. Instantaneous parameter analysis is compared with Fourier analysis and a technique for marking the transitions between certain types of phoneme based on the pdf of instantaneous frequency is developed.

Chapter 5 is concerned with reconstruction of signals from their original or modified instantaneous parameters. Reconstruction distortion caused by poor calculation or storage of instantaneous frequency is illustrated. Frequency shifted reconstruction is proposed as a tool for reconstruction distortion analysis and is used to classify the types of distortion caused by constant amplitude reconstruction of speech. The results of this analysis are extended to infinite peak clipped speech. Frequency division and multiplication of vowels and sub-bands of vowels is illustrated and confirmed to be of little practical use. The remainder of the chapter is then devoted to analysis of lowpass filtered instantaneous parameter reconstructions of various phonemes and recordings of speech. The dependence of speech intelligibility on each of the instantaneous parameters is investigated as are the distortions peculiar to this type of reconstruction.

Chapter 6 consists of a section detailing avenues for future research, followed by conclusions.

A paper describing a method for reduction of the peak/rms power ratio of speech based on analytic signal zero manipulation is currently under preparation. This will be followed by a more general paper on the instantaneous parameters and zeros of speech.

GLOSSARY

(1) DEFINITIONS OF SYMBOLS

$e(t)*h(t)$	convolution of e(t) and h(t)
$\hat{s}(t)$	Hilbert Transform of s(t)
$\overset{FT}{\leftrightarrow}$	Fourier Transform pair
$\Psi^*(z)$	complex conjugate of $\Psi(z)$
$a(t)$	instantaneous amplitude
$\omega_i(t)$	instantaneous frequency
$\overline{\omega_i(t)}$	time average of $\omega_i(t)$
$\Delta\omega$	equivalent rectangular bandwidth
$s'(t)$	differentiation of s(t) with respect to time
$HT\{\ \}$	Hilbert Transform
$sgn(\ )$	signum function
$\delta(t)$	Dirac delta function
$\lvert\ \rvert$	magnitude

(2) ABBREVIATIONS AND TERMS

pdf	probability density function
LHP	lower half of the complex time plane
UHP	upper half of the complex time plane
CCITT	International Telegraph and Telephone Consultative Committee
FFT	fast Fourier Transform
FIR	finite impulse response
TAD	tapped analogue delay line
A/D	analogue to digital
D/A	digital to analogue
REMOVED	refers to analytic signal zero with imaginary component at $+\infty$

CHAPTER 1

(1.1) THE VOCAL APPARATUS

The vocal tract in man is a non-uniform acoustic tube. At one end, the tube is terminated by the glottis and at the other end by the lips or nostrils.

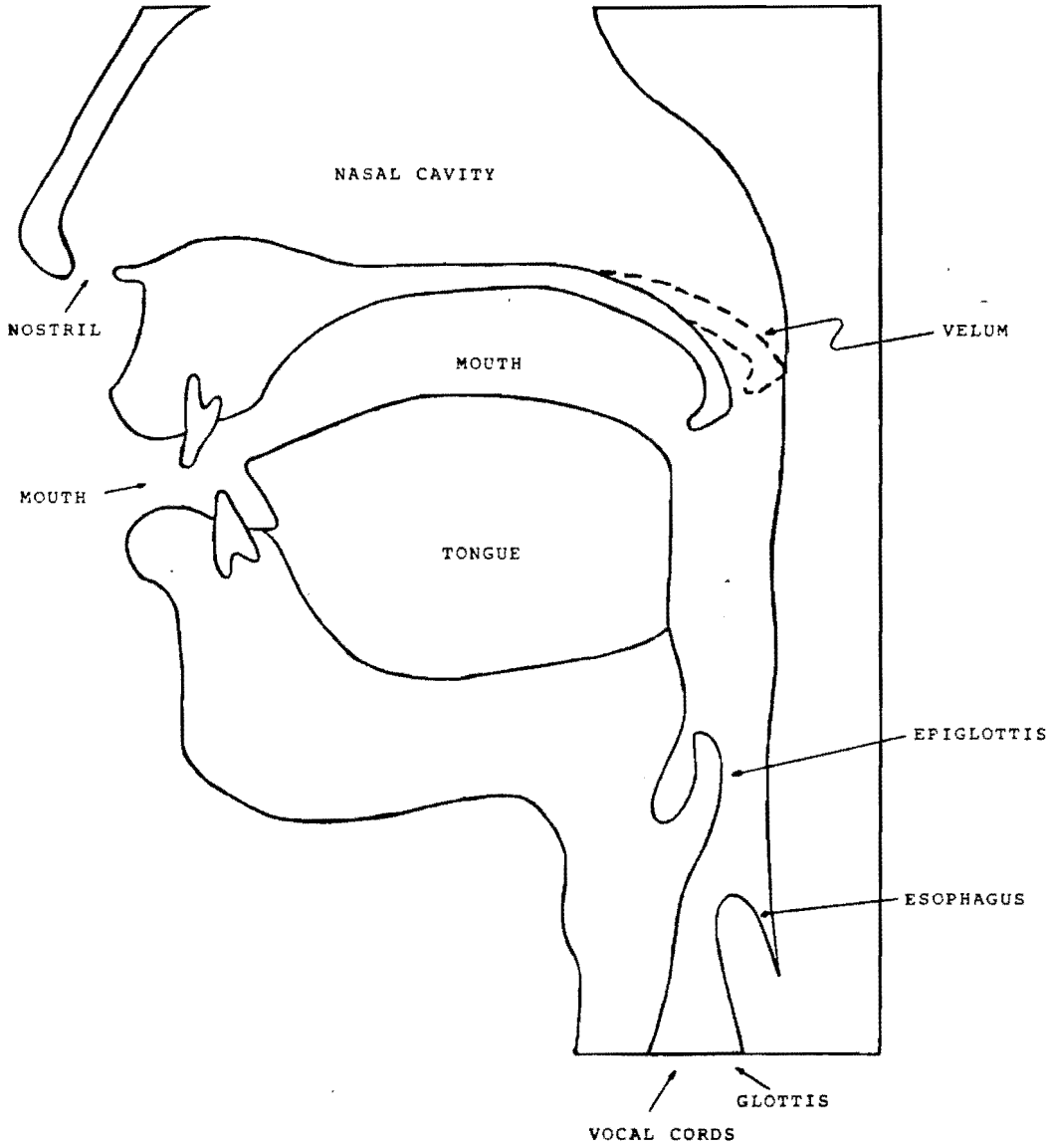


Fig. 1.1 Schematic of Human Vocal Tract

The shape and size of resonant cavities in the tube can be altered by movement of the articulators (jaw, tongue, teeth, lips, and velum). This has the effect of altering the overall frequency response. The nasal tract, which runs from the velum to the nostrils, is a secondary, branching, acoustic tube which is fully closed off or coupled to the vocal tract by movement of the velum (Ref. 1, 2).

Continuous excitation of the vocal tract can be of two kinds; voiced or unvoiced. Both types of excitation involve a flow of air from the lungs, through the trachea and into the vocal tract.

Voiced sounds are produced when the vocal cords, situated at the glottis, are allowed to vibrate in the flow of air. The interrupted flow causes quasi-periodic pressure pulses with a repetition rate usually between 100 and 250 Hz.

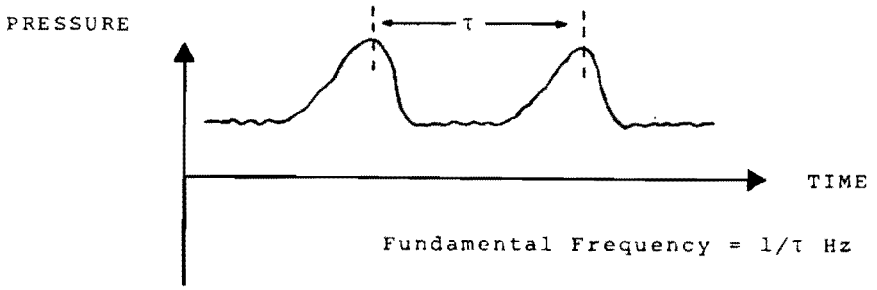


Fig. 1.2 Glottal Pressure Waveform

These pulses are normally assumed to be periodic and spectral analysis reveals a regular line amplitude spectrum with fundamental frequency at the pulse repetition rate. Harmonic amplitude decreases with frequency at approximately 10 dB per octave (Ref. 3, 4).

Unvoiced sounds are produced when the vocal cords are prevented from vibrating and the air flow is forced through a constriction in the vocal tract. If the rate of flow is such that turbulence occurs, the vocal tract is excited by a noise waveform which is spectrally continuous and flat and whose source is at the point of constriction.

Impulsive vocal tract excitation is caused by the sudden release of a pressure block in the tract. Sounds produced in this way are called plosives and are characterised by the position of the air block and whether the vocal cords vibrate during pressure build-up and immediately after pressure release.

Although interactions between the vocal tract and excitation source are difficult to quantify, it is often assumed that the sound source and vocal tract are independent and separable. This leads to the source-filter model figure 1.3.

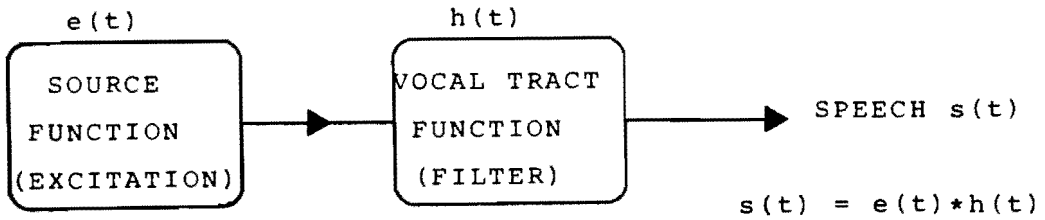


Fig. 1.3 Source-Filter Model of Speech Production

The resulting speech waveform is the convolution of a particular source function and filter function (Ref. 5).

(1.2) THE SPEECH SIGNAL

Speech is most easily represented as a string of acoustic information elements or "phonemes". A particular language is constructed from a finite number of phonemes and these can be categorised by their acoustic properties. The following is a summary of English phonemes grouped acoustically.

(1.2.1) VOWELS

These voiced sounds are produced by the excitation of a stationary vocal tract with glottal pressure pulses. As the nasal tract is fully closed off, the vowel sounds have the frequency response of a fixed vocal tract setting imposed on the line spectrum of the excitation. The tract frequency response usually exhibits three or four major resonances or "formants" in the frequency range 300 Hz to 3400 Hz and each vowel sound lasts for between 20 ms and 500 ms. The amplitude spectrum of a vowel is illustrated in figure 1.4. (Ref 1, 6, 7).

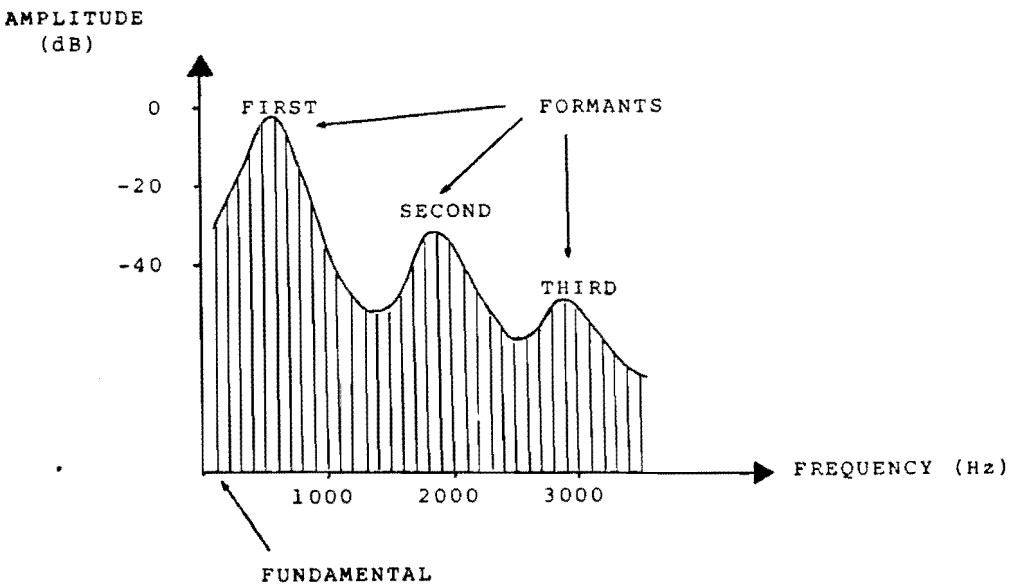


Fig. 1.4 Vowel Amplitude Spectrum

Table 1.1 is a list of the English vowels including key words, formant frequencies and relative formant amplitudes.

VOWEL	KEY WORD	FORMANT FREQUENCY (Hz) (RELATIVE AMPLITUDE (dB)) f_1	f_2	f_3
/i/	beat	270 (0)	2290 (-20)	3010 (-24)
/I/	bit	390 (0)	1990 (-20)	2550 (-24)
/ɛ/	bet	530 (0)	1840 (-15)	2480 (-22)
/æ/	bat	660 (0)	1720 (-11)	2410 (-21)
/ʌ/	but	640 (0)	1190 (-9)	2390 (-26)
/a/	barn	730 (0)	1090 (-4)	2440 (-27)
/ɔ/	bought	570 (0)	840 (-7)	2410 (-34)
/U/	book	440 (0)	1020 (-11)	2240 (-33)
/µ/	boot	300 (0)	870 (-16)	2240 (-40)
/3/	bird	500 Relative formant levels not well defined.	1500	2500

Table 1.1

The information tabulated above is an average based on data from many male speakers. It is possible that an individual's resonant frequencies could vary considerably from those listed.

(1.2.2) CONSONANTS

In general, consonants differ from vowels in that they are not always voiced, may require nasal tract coupling, may involve greater tract constrictions and may depend on vocal tract dynamics.

(1.2.2.1) WHISPERED VOWELS

The whisper /h/ is produced when the vocal tract is excited by air flowing through a constriction at the glottis. In this case, the vocal cords do not vibrate and the nasal tract is closed. For most uses, therefore, the whisper assumes the formant structure, but not the line spectrum, of the vowel it preceeds. Examples of whispers are in the words "help" and "hair".

(1.2.2.2) NASAL CONSONANTS

Nasal consonants are produced by closing the vocal tract near the front (using the tongue or lips) and allowing the nasal tract to carry voiced energy. This results in a nostril radiated sound which exhibits a low frequency resonance and a spectral null at approximately 600 Hz. The remainder of the spectrum is of low amplitude and is almost uniform from 1000 Hz to 2300 Hz. Table 1.2 is a list of the nasal consonants and key words.

Consonant	Key Word
/m/	me
/n/	no
/y/	sing

Table 1.2

Nasal coupling during vowel formation can lead to gross distortion of the vowel formant structure. This is especially noticeable when the important first formant coincides with the nasal spectral null (Ref. 8, 9).

(1.2.2.3) FRICATIVES

Fricatives are generated by turbulent air flow at a tight constriction in the vocal tract. The constricted flow may be unmodulated (unvoiced fricatives) or may have been previously modulated by vocal cord vibration (voiced fricatives).

(1.2.2.3.1) UNVOICED FRICATIVES

These noise-like sounds generally involve less acoustic power than vowels and, unlike vowels, most of their energy is concentrated above 3000 Hz. Table 1.3 is a list of the unvoiced fricatives and key words.

Consonant	Key Word
/f/	for
/θ/	thin
/s/	see
/ʃ/	she

Table 1.3

Each sound is characterised by its corresponding position of constriction. As this may occur part way along the vocal tract, resonances are produced by the portion of the tract after the constriction and anti-resonances by the section before (Ref. 6, 10, 11).

(1.2.2.3.2) VOICED FRICATIVES

These are essentially the vocal tract constrictions associated with unvoiced fricatives, but with a voiced energy source. Table 1.4 lists the voiced fricatives and key words.

Consonant	Key Word
/V/	vote
/ŧ/	then
/z/	zoo
/3/	azure

Table 1.4

Normally, low frequency voiced energy is dominant and the result is a modified formant spectral structure.

(1.2.2.4) PLOSIVES OR STOP CONSONANTS

Generation of plosive sounds relies on vocal tract motion. In general, a point in the oral cavity of the vocal tract is quickly closed or opened causing rapid build-up and release of air pressure. The characteristic interrupted air flow of plosive sounds may be unvoiced or voiced.

(1.2.2.4.1) UNVOICED PLOSIVES

Throughout the entire pressure build-up and release cycle, the vocal cords do not vibrate. During the build-up phase there is a period of silence. The following pressure release produces a transient sound with a short time amplitude spectrum similar to that of an unvoiced fricative.

If, however, the unvoiced plosive is followed by a vowel, there may be a long period of unvoiced aspiration (approximately 50 ms) during which the continuous spectrum

adopts the vowel formant structure.

Table 1.4 is a list of the unvoiced plosives and key words.

Consonant	Key Word
/P/	pay
/t/	to
/k/	key

Table 1.5

(1.2.2.4.2) VOICED PLOSIVES

In this case, there may be voicing during the period of pressure build-up. The resulting sound is not radiated by the mouth or nostrils and subsequently is of low energy.

The pressure release transient is similar to that of an unvoiced plosive. This changes quickly to the formant structure of the following vowel. Table 1.6 lists the voiced plosives and key words.

Consonant	Key Word
/b/	be
/d/	day
/g/	go

Table 1.6

The points oforal cavity closure for corresponding voiced and unvoiced plosives are identical. Voiced plosives, however, usually require a greater pressure build-up (Ref. 12).

(1.2.2.5) GLIDES AND SEMIVOWELS

Glides are dynamic sounds which normally preceed a vowel. They possess a formant structure which rapidly changes to that of the following vowel.

Semivowels are stationary sounds, but differ from vowels in that the vocal tract may be more constricted (the tongue not fully down). This leads to differences from vowel formant structure.

(1.2.2.6) DIPTHONGS AND AFFRICATIVES

Some of the previously described phonemes may be combined to form dipthongs and affricatives. These sounds are wholly dependent on vocal tract dynamics.

Dipthongs are defined as "mono-syllabic items of speech" which start at one vowel sound and move toward another. /eI/ as in "say" is an example.

Affricatives are changes from some plosive sounds to fricatives. /tʃ/ in "chew" is an example (Ref. 1).

(1.3) HEARING AND PERCEPTION

Ignoring the transmission medium, hearing and perception by the human brain in the next aspect of the speech communication system to be considered. To determine the information bearing qualities of speech, it is necessary to define the capabilities and limitations of the ear - brain receiver. This is a very difficult task as the processes involved are only partially understood.

(1.3.1) PHYSIOLOGY OF THE EAR

Figure 1.5 is a simplified illustration of the anatomy of the human ear.

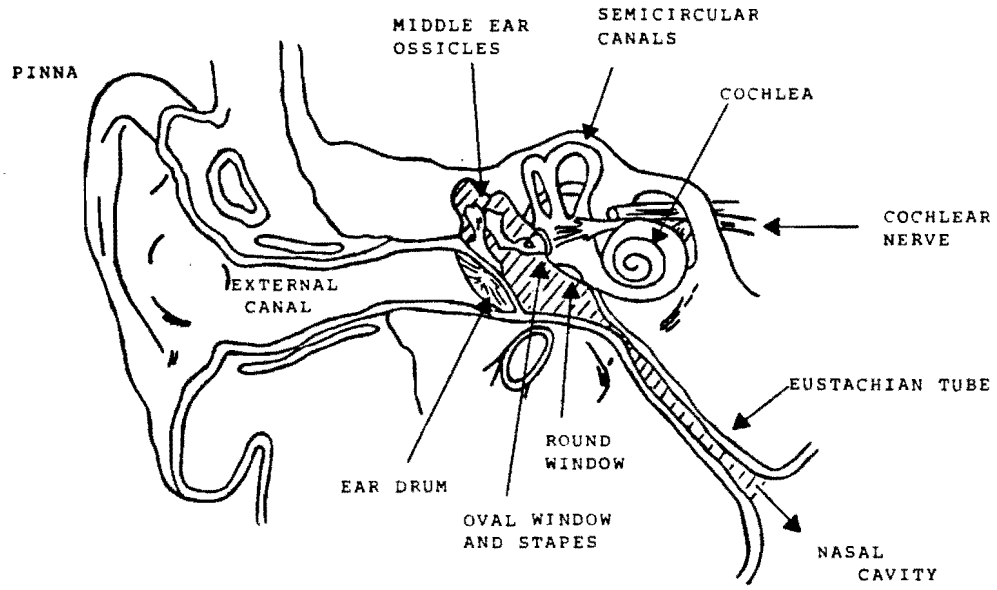


Fig. 1.5 Schematic of Human Ear

Sounds (speech) from a free-field source reach the inner mechanisms of the ear via the pinna and external canal. The complex physical structure of these "outer ear" components cause frequency dependent diffraction effects. These are used in the discrimination of sound direction when a source is equidistant from both eardrums. The total frequency range of normal human hearing is approximately 15 to 25000 Hz. (Ref. 13, 14).

The modified sounds impinging on the eardrum are transmitted to the inner ear via a lever arrangement of three small bones. These bones are enclosed within the middle ear cavity and are illustrated more clearly in figure 1.6. The middle ear transmits vibrations to the oval window of the cochlea.

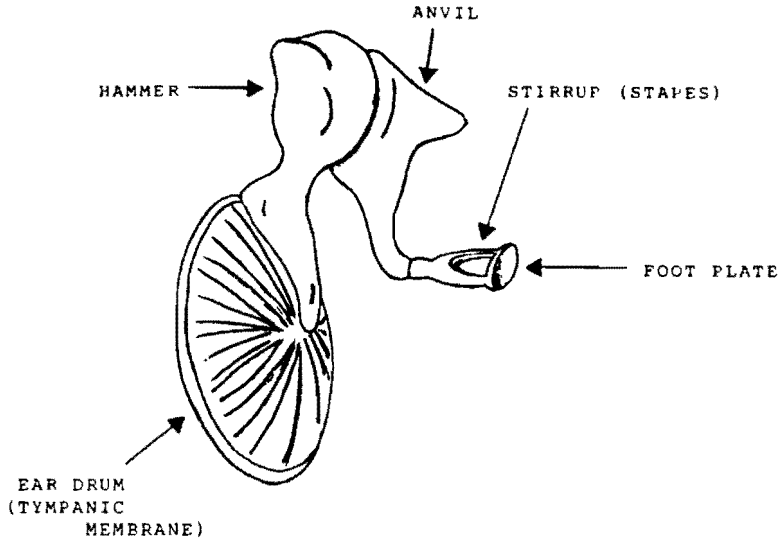


Fig. 1.6 Middle Ear

Taking into account the diameters of the eardrum and stapes and the lever ratio of the middle ear bones, the middle ear acts as an impedance transformer between the low impedance air and the fluid filled cochlea. Although the observed transformer action greatly increases the efficiency of power transfer, the transformer ratio is far from optimal. The process is also frequency dependent (Ref. 13).

The inner ear is by far the most complex anatomical structure of the ear. Although it contains both the semi-circular canals and the cochlea, it is only the cochlea that is concerned with auditory processing. Figure 1.7 is a schematic drawing of a straightened cochlea.

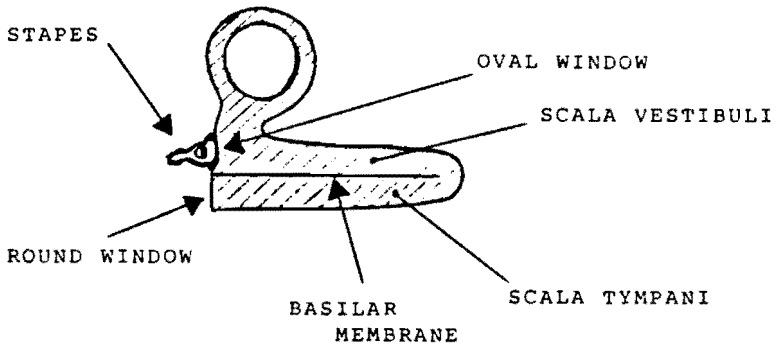


Fig. 1.7 Schematic of Cochlea

The stapes vibrating against the oval window has the effect of setting up waves in the cochlear fluid. These waves travel through the inner chamber (scala vestibuli) reaching the outer chamber (scala tympani) by way of the cochlear partition or by curving around the tip of the spiral. The waves terminate at the round window (Ref. 15).

Innervation of the cochlea is both afferent and efferent implying a complex neural control system and an unusual method of information coding at the frequency selective basilar membrane.

Many attempts have been made to explain the ears observed frequency selectivity. Helmholtz envisaged a succession of tuned strings (or contiguous band-pass filters), but the ears observed speed of response is not consistent with a filter system which can detect a frequency change of 3 Hz at 1 kHz (Ref. 13).

Physical examination of the basilar membrane lead von Békésy to a theory of travelling waves in the cochlea (Ref. 16). The theory is essentially that travelling waves set up in the cochlear fluid correspond to frequencies present in the sound source. A travelling wave causes a maximum average displacement of the basilar membrane at a given point along its length for a given frequency. The fact that high frequency travelling waves were quickly attenuated and low frequency waves travelled further gave the basilar membrane the appearance of a non-uniform transmission line.

Unfortunately this mechanical filter is still not sufficient to explain the frequency selectivity of the ear. Any further "sharpening" of the filter process must occur during or after the coding of the audio signals into neural signals.

Neural encoding of sounds is performed by the hair cells. These are mounted on the "organ of corti" which is perched on the basilar membrane. Figure 1.8 is a cross-section of the cochlea illustrating cochlear innervation.

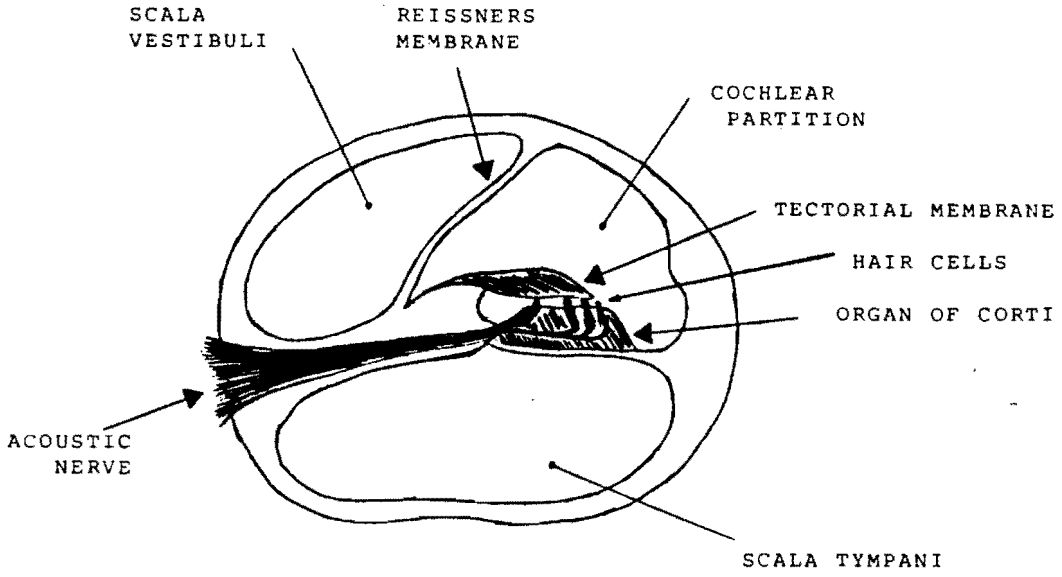


Fig. 1.8 Cochlear Cross Section

The tufts of hair growing on each hair cell are believed to be attached to the "tectorial membrane". Displacement of the basilar membrane causes deformation of the hairs which results in a flow of ionic current and eventual firing of afferent nerve fibres.

Some generalisations can be made about neural coding. For frequencies up to 3000 Hz, the evoked response in nerve fibres (all or nothing spikes) appears to follow the frequency of an applied tone (e.g. 500 responses per second for a 500 Hz tone). However, as the refractory period of a nerve fibre is about 1 ms, a single fibre cannot fire at a rate of more than 1000 per second. This leads to the concept of "volleying" or two or three fibres transmitting every second or third spike.

For frequencies above 4000 Hz, coding appears to be wholly dependent on the position of the displacement maximum on the basilar membrane. Loudness information is also coded in the timing and spatial distribution of neural impulses.

(1.3.2) PERCEPTION

Psychophysical measurements indicate that the ear acts as a bank of overlapping filters with bandwiths ranging from around 200 Hz at 1 kHz to 2 kHz at 10 kHz. It has an apparent dynamic range of around 120 dB, but the ears sensitivity is both frequency and signal dependent.

(1.3.2.1) SIMPLE SOUNDS

Any measure of the ears frequncy response must be a function of "reported" signal level and the characteristics of the applied signal. The most straightforward result is obtained as perceived sound pressure level for a pure tone stimulus. The curves for an average listener are plotted in figure 1.9 with relative loudness measured in "phons".

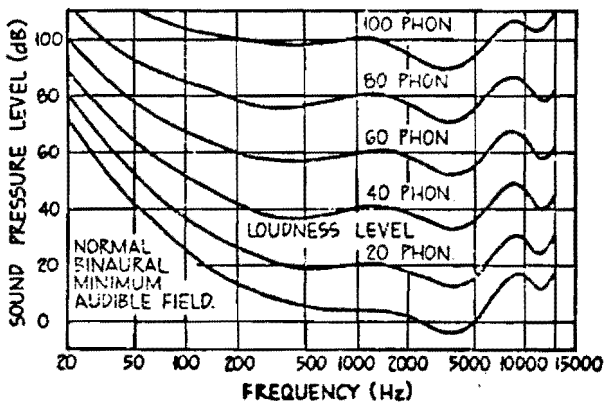


Fig. 1.9 Equal Loudness Contours of Sound Pressure Level

Expanding the range of test stimuli to simple bandlimited sounds allows the measurement of minimum detectable frequency deviations and introduces the concept of masking.

The minimum detectable fluctuations of the frequency of a pure tone is dependent on its starting frequency. In accordance with the overlapping filter model, the ear is more sensitive to frequency deviations at low absolute frequency than it is at high frequency. Results for an "average listener" are tabulated in table 1.7 (Ref. 7).

CENTRE FREQUENCY (Hz)	62.5	125	250	500	1000	2000	4000	8000
DETECTABLE FREQUENCY DEVIATION (Hz)	2.7	3.1	2.9	2.5	3.0	4.6	9.0	29.5

Table 1.7

Masking essentially quantifies the ears ability to detect a signal in noise. If a test tone of sufficient sound pressure is presented to the ear, it will be perceived. If, then, a continuous spectrum bandlimited signal, centre frequency at that of the test tone and the same sound pressure level is presented simultaneously, the original signal may or may not be swamped (masked). The bandwidth of the "noise signal" required to mask the sinusoid is the "critical bandwidth". Critical bandwidth is a function of centre frequency and table 1.8 presents some average values.

CENTRE FREQUENCY (Hz)	500	2000	4000	8000
CRITICAL BANDWIDTH (Hz)	50	100	200	600

Table 1.8

When interpreting the results of figure 1.9 and tables 1.7 and 1.8, the frequency range 300 Hz to 3400 Hz is of

particular interest. This range is known as the "telephone bandwidth" and has been defined to be the minimum bandwidth capable of passing intelligible speech of sufficient quality that the speaker may be easily recognised.

Figure 1.9 shows that the perceived loudness of tones is fairly constant across the telephone bandwidth. Tables 1.7 and 1.8, however, indicate that the ear has better frequency resolution at low frequency and that it is also more prone to masking at low frequency.

(1.3.2.2) COMPLEX SOUNDS

Perception of complex continuous sounds involves recognition of qualities such as pitch and timbre.

Flanagan (Ref. 1) describes pitch as a "subjective attribute" which ranks a sound on the scale of low to high. Being subjective, pitch is difficult to measure physically and because of human ear non-linearities, perceived frequencies may not even correspond to frequencies of signal components.

If the overlapping filter model of hearing were accurate, a continuous complex sound would be fully characterised by its amplitude spectrum. Changing the phase of particular components of a complex sound, however, can cause considerable time waveform changes which the ear may detect as changes of timbre. Once again this effect is subjective and difficult to quantify (Ref. 13, 17, 18).

(1.3.2.3) TRANSIENT SOUNDS

Investigation of the perception of speech must consider the way in which the ear processes transient sounds.

On average, the pitch of a simple sinusoid cannot

be accurately assigned unless it is present for at least 50 ms. More complex sounds may require up to 200 ms for the pitch, and therefore the sound to be fully perceived. Conversational vowels usually persist long enough for pitch perception (Ref. 7).

(1.4) CONVERSATIONAL SPEECH

Conversation, or communication by speech, is a combination of speech production and perception. An utterance of speech is produced by stringing together a set of words, from a fixed vocabulary, within the grammatical rules of the language. Although this conveys a basic message, further information is imparted in the form of speaker recognisability, emotional content and personal information such as the speakers background from accent and choice of words (Ref. 19).

Word recognition involves both the auditory capabilities of the ear and correlative capabilities of higher centres of the brain. Superimposed information, such as emotional content, may be inferred from unusual pitch variation, but experience and correlative memory are always required.

There are many current theories which partially explain how the human brain recognises a particular phoneme without being confused by the idiosyncrocies of individual speakers. Models have been proposed and these often form the basis of speech recognition algorithms.

(1.4.1) INFORMATION CONTENT

It has been estimated that approximately 10 phonemes are uttered per second during conversational speech. This corresponds to transmission of the equivalent written

information at a rate of less than 50 bits per second. (Ref. 1, 7).

Shannon's equation for the maximum rate of error free data transmission over a continuous channel is

$$C = B.\log_2\{(S/N)+1\} \quad \ldots (1.1)$$

where C is the channel capacity in bits per second, B is the bandwidth of the channel and (S/N) is the "signal to noise ratio" (Ref. 20). For the case of a 3000 Hz bandwidth telephone voice channel with signal to noise ratio of 30 dB, equation 1.1 predicts a possible information rate of around 30,000 bits per second.

Although the 50 bits per second based on phoneme rate does not take personal or emotional information into account, there is still a striking difference between the channel capacities calculated by the two methods. The ratio of 50 : 30,000 implies a possible information rate reduction by a factor of almost 1000.

This type of calculation illustrates the basic trade-off in efficient speech transmission. High fidelity voice carries intelligibility plus all the additional personal information. Telephone quality voice maintains intelligibility and has enough personal content for the speaker to be recognisable and reasonably pleasant to listen to. Any further reduction of information transmission rate (corresponding to bandwidth reduction for fixed signal to noise ratio) can cause audible signal degradation and if a special coding scheme has not been used, may reduce intelligibility.

Intelligible speech for which the information associated with acoustic fidelity has been lost or scrambled is often described as being of "communications quality".

(1.5) SPEECH CODING

For transmission over a man-made channel, speech is usually converted from an acoustic pressure wave to an electrical or optical signal. In a standard baseband telephone (300 to 3400 Hz bandwidth), the transmitted signal is a direct electrical analogue of the pressure wave.

With the current demand for additional and more reliable speech transmission channels, a great deal of research has been carried out in the field of elaborate speech coding algorithms. Some techniques provide a more robust signal for transmission but many are concerned with reducing the bandwidth or bit rate required to transmit acceptable quality speech. These bandwidth reducing schemes are designed around some redundancy or redundancies of the speech signal.

Due to the large volume of literature on the subjects of speech coding and bandwidth reduction, what follows is merely an overview. It is hoped, however, to cover the essential topics.

(1.5.1) VOCODERS

The first low bandwidth speech coders were developed before the introduction of reliable digital hardware. For this reason, the algorithms were suitable for analogue implementation and the transmitted signals were also analogue in nature. These early vocoders (voice-coders) were based on the excitation - vocal tract model of speech production. (figure 1.3).

The philosophy behind Dudley's original Channel Vocoder (Ref. 21) was essentially that transmission of the acoustic speech waveform is unnecessary and that slowly

varying excitation and vocal tract parameters may be transmitted instead. This leads to the concept of "analysis - synthesis" telephony.

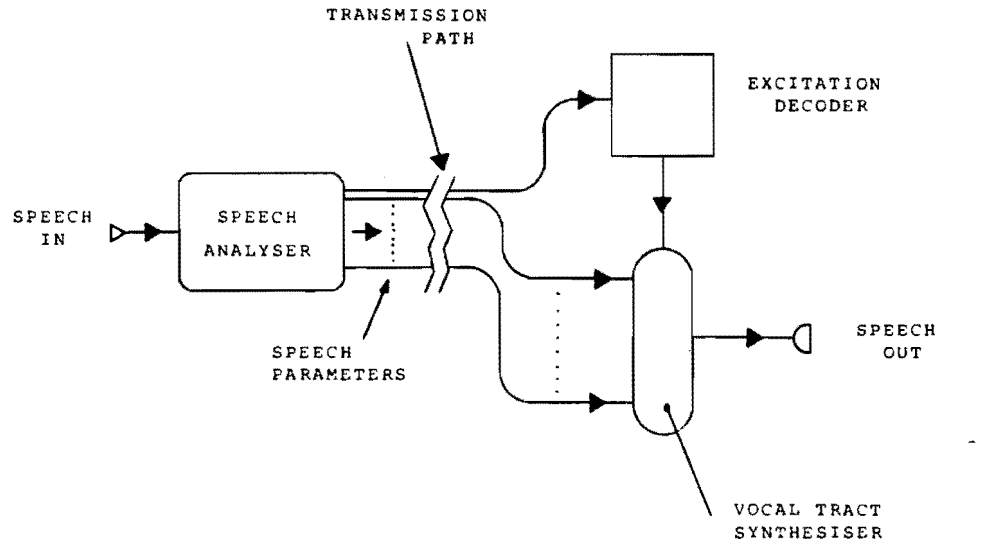


Fig. 1.10 Analysis - Synthesis Telephony

The channel vocoder transmits three types of data. They are a decision on whether the speech is voiced or unvoiced, an estimate of the rate of excitation pulses (if voiced) and a representation of the vocal tract filter function. Figure 1.11 is a block diagram illustration of the channel vocoder.

When voicing is detected, the glottal pulse rate (or fundamental frequency) is extracted and is coded for transmission to the receiver's pulse source as a low pass voltage. The pulse source produces a waveform which attempts to reproduce the wave shape of voiced glottal pulses.

If no voicing is detected, the excitation selector switch in the receiver toggles to "noise source". The

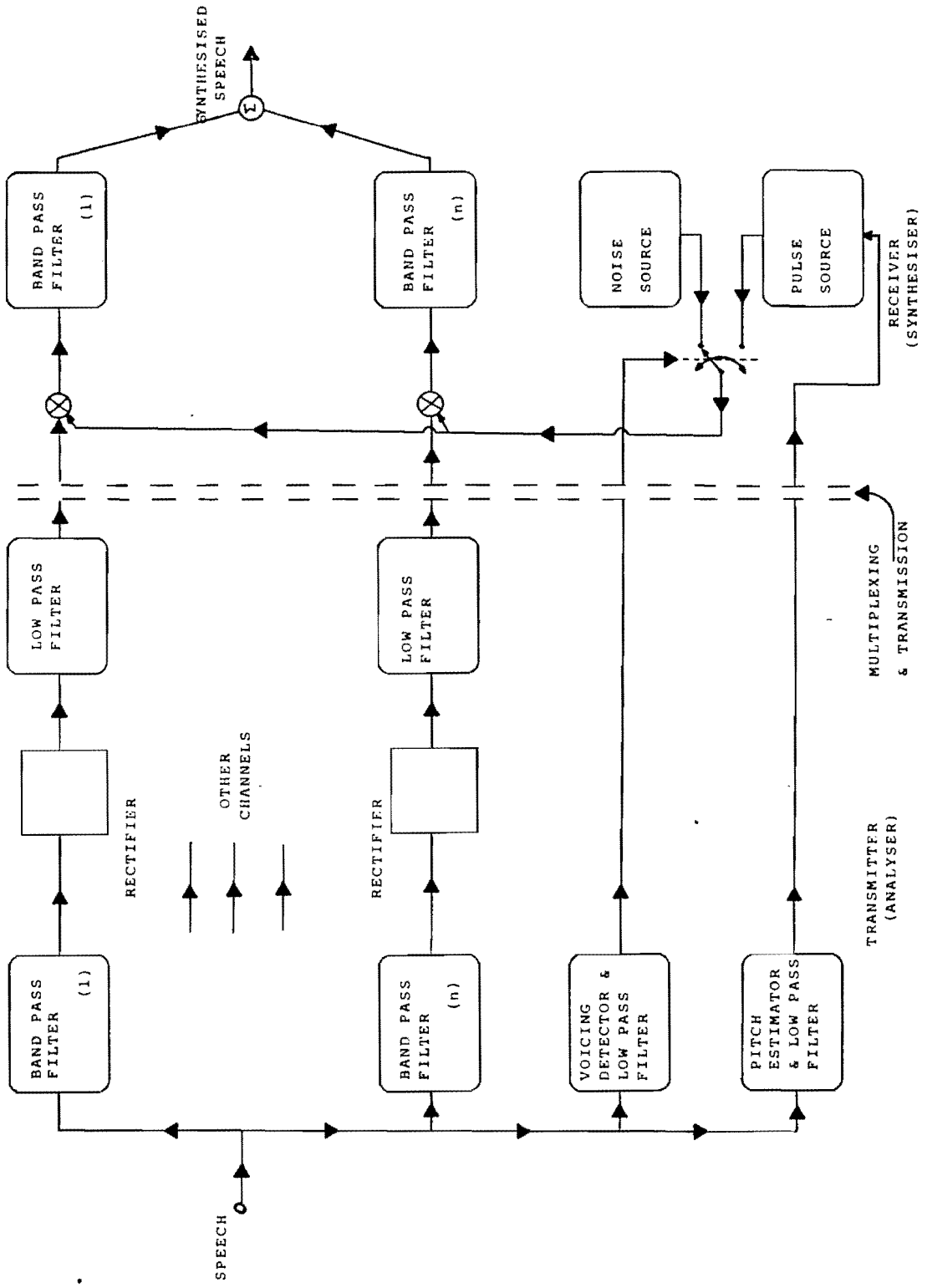


Fig. 1.11 Schematic of Channel Vocoder

excitation amplitude spectrum is now flat and continuous and can be used for synthesising unvoiced sounds.

The vocal tract transmission function is estimated by the transmitter with a set of (about 10) equal bandwidth, contiguous bandpass filters. The filters combine to cover the entire baseband telephone bandwidth. The output of each filter is rectified and low pass filtered resulting in each channel transmitting an estimate of the spectral power at the particular band centre. Taken as a whole, the low pass filter outputs are a representation of the signals short time power spectrum.

As indicated previously, certain phonemes can be considered stationary over a period of tens of milliseconds. Ideally then, the power spectrum, and thus the vocoder channel signals, will vary slowly, except at the transition from one phoneme to another.

At the receiver (synthesiser) spectral information is decoded by using each channel signal to amplitude modulate the excitation and passing the resulting signals through the appropriate channel of another set of bandpass filters. The receiver bandpass filters should be identical to those in the transmitter.

This type of vocoder claims an analogue bandwidth compression capability of about 10 to 1, but intelligibility of the synthesised speech is speaker dependent (Ref. 22). Particular types of distortion introduced by vocoders will be discussed later.

An early refinement of the channel vocoder was prompted by the discovery that there is often a high degree of correlation between channel signals and that some channels carry no signal for significant periods of time (Ref. 23).

This lead to the idea of the "peak picker" algorithm.

The peak picker is essentially a channel vocoder analysis system, but it is one in which only channels carrying significant energy are transmitted. The number of channels transmitted is always fewer than the total number of analyser channels, resulting in greater bandwidth savings (Ref. 24).

Another step in the refinement of channel transmission is to extract only the frequency and amplitude of vocal tract resonances. As there are up to three major resonances in the frequency range 300 to 3400 Hz, only three amplitude and three frequency channels need be transmitted along with the normal voicing and excitation information. The resulting device is the Formant vocoder. It requires less bandwidth than the channel and peak picker vocoders but performs better with voiced phonemes than with unvoiced fricatives (Ref. 25, 26).

The Phonetic Pattern Recognition Vocoder obtains even greater bandwidth savings. This device examines the spectral properties of each phoneme and compares them with a stored "vocabulary". When a match is made, a code representing the particular stored phoneme is transmitted along with voicing and pitch information.

At the receiver, the code is matched to the phoneme by means of an identical vocabulary and the excitation shaped by the particular phoneme filter response. Intelligibility of the resulting synthesised speech is speaker dependent and there is little possibility of recognising the speaker. The overall transmission bandwidth may be as low as 100 Hz (Ref. 27).

The introduction of sophisticated technology and sampled data systems brought about a revolution in analysis/synthesis

telephony. In particular the ability to design and build recursive digital filters made analysis of the speech waveform by linear prediction possible. As there are many algorithms based around this theme, only the basic principles will be covered in this overview.

Figure 1.12 is the diagram of an analysis - synthesis telephony system based on linear prediction of the speech waveform.

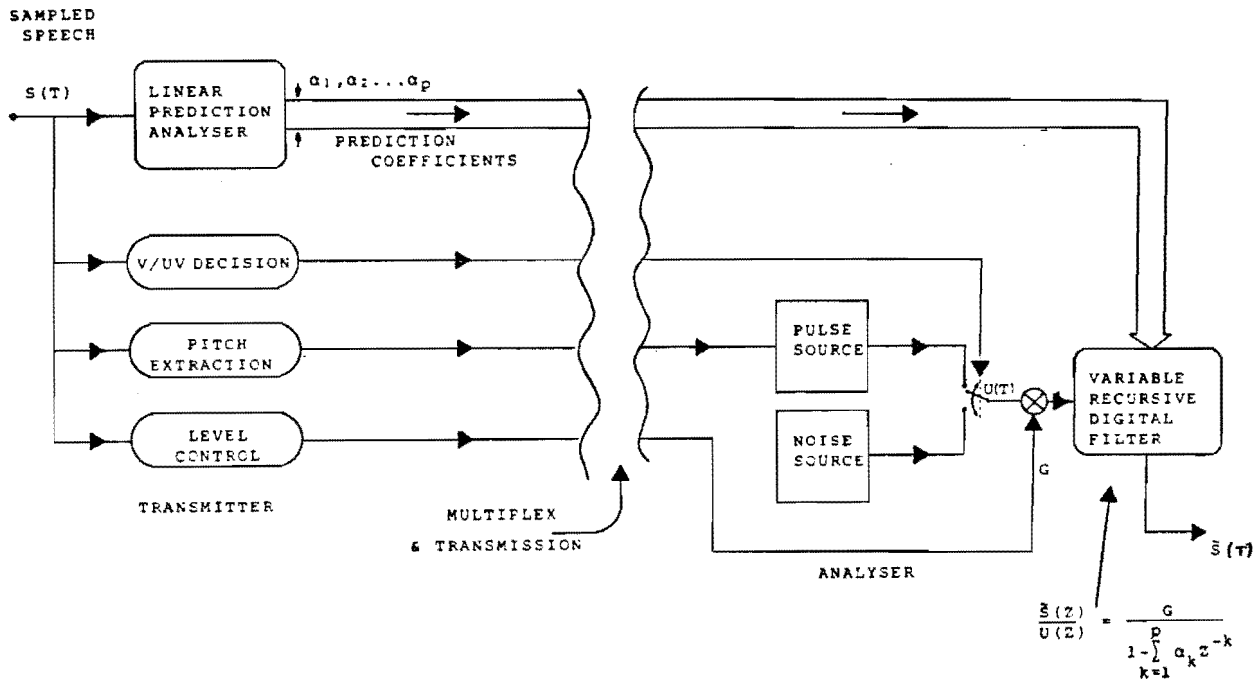


Fig. 1.12 Linear Prediction Telephony System

The transmitter must send the pitch period, a voiced/unvoiced decision, an estimate of the signal energy and the set of prediction parameters α_1 to α_p. By comparing this with the channel vocoder, it can be seen that the prediction

parameters are equivalent in information content to the bandpass channel signals (vocal tract filter function).

The receiver is an implementation of the source-filter model of speech production (figure 1.3). In this case, the vocal tract filter transfer function

$$\frac{\tilde{S}(Z)}{U(Z)} = \frac{G}{1-\sum_{k=1}^{p} \alpha_k Z^{-k}} \qquad \ldots (1.2)$$

is an "all pole" version. If the order of the filter, p, is high enough this transfer function is capable of approximating both the resonances and antiresonances found in the vocal tract frequency response of band limited consonants and vowels.

Prediction coefficients are calculated on the basis of minimising the average squared prediction error E, over a period of N predictions, where prediction error e(T) is defined

$$e(T) = S(T)-\tilde{S}(T) \qquad \ldots (1.3)$$

and

$$E = \sum_{m=0}^{N} e^2(T+m) \qquad \ldots (1.4)$$

The minimisation is achieved by setting the partial differentials of E with respect to each αi to zero

$$\frac{\delta E}{\delta \alpha i} = 0 \ , \ 1<i<p \qquad \ldots (1.5)$$

and solving the subsequent set of simultaneous equations. The autocorrelation and covariance methods are common techniques for solving these equations.

Linear prediction parameters are easy to transmit as they are very slowly varying. For this reason, they also lend themselves to speech storage application. The major advantage of linear prediction over the channel vocoder is the ability to dispense with banks of analogue filters. (Ref. 28, 29, 30, 31, 32).

(1.5.1.1) VOCODER PERFORMANCE

Factors which influence the performance of the above types of vocoders are:

1. The importance of the transmitted parameters to intelligibility.
2. The importance of the transmitted parameters to recognisability.
3. How well the parameters are estimated.

The importance of a parameter to the intelligibility and recognisability of synthesised speech may be inferred from knowledge of the speech production and perception mechanisms. Parameter extraction methods are often designed intuitively but are usually based on knowledge of speech production.

Transmitted parameters common to all of the above vocoders are an estimate of pitch and a voiced/unvoiced decision.

(1.5.1.2) PITCH ESTIMATION

Efficient pitch estimation is important as it is known that a good approximation to the pitch contour of voice facilitates speaker recognition. Existing pitch estimation algorithms are best considered under the three classifications; time domain estimators, frequency domain estimators, and hybrid estimators. (Ref. 33, 34)

Time domain estimators operate on the temporal speech signal. Waveform features which can be used include pulse-like variations of amplitude or energy, time differences between zero-crossings and features of the waveforms autocorrelation function (Ref. 35, 36, 37, 38).

Frequency domain estimators make use of the fact that periodic excitation imposes a line spectrum "fine structure" on the short time amplitude spectrum of a voiced phoneme. The most simple frequency domain estimator is therefore a filter tracking the fundamental component or line. Unfortunately, such a filter may not be able to follow rapid pitch variations and would be influenced by the large spectral peak of the first formant (Ref. 39).

Cepstral analysis is probably the most elegant, but a very complex frequency domain pitch estimator. Essentially, the cepstrum is the spectrum of the logarithmic amplitude spectrum of speech. In this form, the voiced line structure is represented as a high "quefrency" spike-like component and is easily separated from the low quefrencies of the spectral envelope. The process is analogous to deconvolution of the excitation and vocal tract functions in the time domain (Ref. 40).

Hybrid detectors make use of both the time and frequency domains. For example, frequency domain methods may be used to spectrally flatten a signal which is then analysed by a time domain estimation procedure (Ref. 41).

(1.5.1.3) VOICED/UNVOICED DETECTION

Many early vocoders used the pitch detector as a voiced/unvoiced decision maker. Significant energy at the output of the pitch detector implied a voiced sound and the decision was made accordingly. This strategy encounters difficulties however, when presented with an ambiguous

phoneme, such as a voiced fricative. In this case, the voiced energy may not be sufficient to trigger a voiced decision.

The modern approach to this problem is to apply pattern recognition techniques by monitoring several of the temporal and frequency domain characteristics of the speech signal. Reliable voiced/unvoiced classification is thus possible for speech segments as short as 10ms (Ref. 42, 43).

(1.5.1.4) VOICE EXCITATION

An elegant method of obtaining pitch and a voiced/unvoiced decision without pitch detection is to transmit a sub-band of the speech signal (say 250 Hz to 1 kHz) (Ref. 1). The remainder of the speech bandwidth is coded by a vocoder method.

At the receiver, the sub-band is used directly as the low frequency component of the synthesised speech (250 Hz to 1 kHz) and indirectly as excitation for the coded upper frequencies. The resulting synthetic speech is of improved quality but there is a penalty of increased transmission bandwidth.

(1.5.1.5) VOCODER SPEECH QUALITY

It has been proposed that there is an insurmountable speech quality limitation inherent in low transmission rate vocoders. This is well illustrated by the channel vocoder which, with sufficient channels to represent the speech spectral envelope, still tends to exhibit a "machine accent" or "flatness". Such distortion is often caused by errors in following pitch contours or in voiced/unvoiced decision making, but it should be noted that the separable excitation - vocal tract response model on which these vocoders are based, is only an approximation of the physical situation.

Another assumption inherent in many vocoders is that an accurate description of the short time amplitude spectrum only is required to re-synthesise speech. It is known, however, that altering the phase spectrum can cause significant changes to the speech waveform. Although usually intelligible, phase distortion may cause speech to become "buzzy" or "muffled" (Ref. 13, 44).

Several attempts have been made to classify the quality of synthesised speech and a summary of qualities versus transmission rates will be presented later in this chapter.

(1.5.2) WAVEFORM CODERS

Waveform coding schemes have come about principally as solutions to the problem of representing speech in digital form. Depending on their complexity and intended use, some techniques make use of known statistics or redundancies inherent in time series descriptions of speech (sampled speech).

(1.5.2.1) P.C.M. BASED SYSTEMS

Speech, which has been bandlimited to a telephone voice channel (300 Hz to 3400 Hz), may be uniquely defined by waveform amplitude samples taken at a rate of 8000 per second. This is equivalent to time discretisation of the continuous waveform and leads directly to a pulse amplitude modulation (P.A.M.) representation of the signal. Before they are represented digitally, each sample must also be quantised in the amplitude domain. The result of this second discretisation is that the speech becomes a string of digital words. In this format a waveform may be transmitted as a pulse code modulated signal (P.C.M.).

Even in a coding system as conceptually simple as P.C.M., waveform statistics may be used to improve quantising

efficiency. Speech exhibits a non-uniform amplitude distribution in which there is a greater probability of low sample values. The ideal speech quantiser, therefore, exhibits small quantisation steps at low amplitude and larger steps at high amplitude.

Such non-linear biasing is a common precursor to digital speech transmission and is usually performed by a "compander". A companding system linearises the speech amplitude distribution prior to linear quantisation and transmission.

On reception, the second component of the companding system biases the sample amplitudes back to their original distribution. The resulting compressor-expander system can lead to bit rate reductions of about 30% compared to conventional P.C.M. (Ref. 45, 46).

An even more suitable quantiser system is one which changes its quantisation step size in response to signal power level. Such adaptive quantisation is very appropriate for use with speech as there can be a power difference of around 40 dB between syllables.

Another useful statistic of time quantised speech is the large correlation factor between adjacent amplitude samples. Systems designed to make use of this type of redundancy normally code only the amplitude difference between successive samples. The original waveform is recovered by accumulating the differences or integrating the received signal.

Differential pulse code modulation (D.P.C.M.) makes direct use of sample to sample correlation by digitising and transmitting the amplitude differences as a string of digital words. The advantage over basic P.C.M. is that one bit can often be dropped from the transmitted word length. (Ref. 45).

Even greater use can be made of the correlation by encoding the amplitude difference between the new sample and the output of a linear predictor (Section (1.5.1)). This form of predictive D.P.C.M. is often combined with adaptive quantisation and called A.D.P.C.M.

A drawback common to all of the above P.C.M. based transmission systems is the need for a reasonably fast analogue to digital converter with between 8 and 12 bits resolution. As these have been expensive in the past, a single bit form of D.P.C.M. was developed called Delta Modulation (D.M.). The basic form of D.M. involves transmitting a digital "1" if the next sample amplitude is greater than the current, or "0" if it is smaller. In this case the analogue to digital conversion is performed by a comparator.

The basic D.M. receiver accumulates the positive steps ("1"s) and negative steps ("0"s) to repoduce a granular version of the original signal.

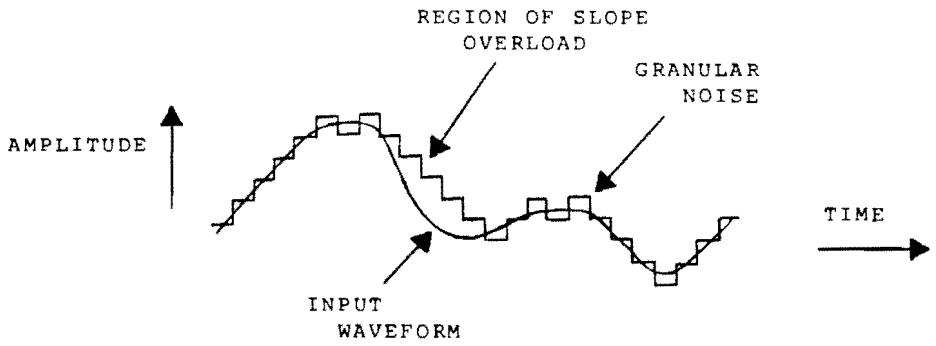


Fig. 1.13 Output of Basic Delta Modulator

The output of the accumulator is low pass filtered before playback.

In order to generate a reasonable facsimile of the input, the D.M. must sample at well above the Nyquist rate (about 8000 Hz for telephone channel). Insufficient sampling rate or insufficient quantisation step size lead to slope overload distortion. This can be rectified by increasing the sampling rate at the cost of increased transmission bandwidth or by increasing step size with the corresponding increase in granular noise (Ref. 45, 47).

In most practical D.M. speech transmission systems, the slope overload versus bandwidth problem is overcome by an adaptive step size algorithm (A.D.M.). The basic step size in an adaptive delta modulator is small to minimise granular noise for a non-varying input. If, however, the comparator output produces a series of several 1's or several 0's, it may be inferred that the linear D.M. is not following the waveform slope and step size is temporarily increased (adapted) until the output "catches" the input. There are many adaption schemes in use today (Ref. 6, 45, 48).

(1.5.2.2) SUB-BAND CODERS

The quality of digitally encoded speech can be improved by coding with respect to perceptually significant signal components. In the frequency domain, the lower frequency components of voice appear to be more important than high frequencies, implying that different coding schemes should be used on each. This leads to the concept of sub-band coding (Ref. 49, 50, 51).

Many sub-band coders split the speech signal into about four fixed or variable sub-bands, each of which is then quantised to a suitable level. Only some waveform coding schemes can then be used to code the sub-bands as the signals do not display a high degree of correlation between adjacent samples.

(1.5.3) FREQUENCY DIVISION VOCODERS

A third class of speech coder, which is of particular interest to the author, is based around the concept of frequency division. These are analysis - synthesis telephony systems but, unlike the vocoders described previously, they do not rely on the excitation - vocal tract model of speech production. They are not classed as waveform coders, however, as they do not operate solely on the speech time waveform.

One of the earliest practical frequency division coders was the VOBANC (Ref. 52). This system bandpass filters the speech signal into three fixed sub-bands, each ideally containing one of the major formant frequencies. Using regenerative modulation techniques, each of the sub-bands is then "divided by two" (thus halving their frequency excursions), and transmitted over half of the original bandwidth. In the receiver, the signals are "multiplied by two" and recombined to form the synthesised speech.

An improved dividing algorithm and better results are claimed for the CODIMEX speech compression system (Ref. 53). This device also separates the speech signal into three sub-bands, but instead of straight frequency division, it performs a "signal rooting" operation. The output of each sub-band filter, S(t), may be represented

$$S(t) = a(t)\cos\phi(t) \qquad \ldots (1.6)$$

where a(t) is defined to be the "instantaneous amplitude" and $\phi(t)$ the "instantaneous phase". By a reasonably simple modulation process, the system generates a signal which is proportional to the square root of S(t)

$$g(t) = (a(t))^{\frac{1}{2}}\cos(\phi(t)/2) \qquad \ldots (1.7)$$

Three repetitions of the process generates $(S(t))^{1/8}$. The rooted signal is then transmitted over an eighth of the original bandwidth and re-expanded on reception. "Telephone quality" voice is claimed on synthesis by recombination of the three signals.

The "Analytic Signal Rooter" (Ref. 54) operates on the same theoretical basis as the CODIMEX and a computer simulation has indicated good results for a bandwidth reduction of one half. The success of signal rooting is very signal dependent, however, working best on spectra which can be described as the result of large index frequency modulation. The individual vowel formants can be loosely characterised as such.

Assuming that a single formant is suitable for frequency division, there are still several factors limiting the performance of the above systems. Although the normal positions of the three formants can be separated into three sub-bands, there will be some crossover for particular speakers. This implies a finite probability of there being two formants within a sub-band at a particular time and thus unreliable operation of the frequency divider.

Ignoring the amplitude compression effect of signal rooting, frequency division has the effect of dividing the frequency of the dominant spectral component (largest amplitude) and maintaining the frequency spacing of surrounding components (Ref. 55). The bandwidth of the whole signal is not really divided, and passing through a half bandwidth bandpass filter for transmission must introduce distortion. Although this effect was not particularly well investigated for the above systems, it was believed that it might be responsible for a strange "burbling" in the synthesised speech.

The ideal signal for frequency division is an unmodulated sinusoid. As there is no associated bandwidth,

division becomes a frequency translation. Speech can be represented as a set of sinusoid-like signals by prefiltering through a set of contiguous narrowband filters (comb filter). For a vowel, assuming sufficient filters, the filter outputs are fixed frequency sinusoids with slight amplitude modulation. For unvoiced sounds, the outputs are narrowband noise signals which can be approximated by sinusoids.

The Bandpass Compressor (Ref. 56) or Phase Vocoder (Ref. 57) make use of such a set of bandpass filtered signals (usually 30 in the range 300 Hz to 3400 Hz) to perform speech bandwidth reduction and time scale expansion or compression. Figure 1.14 is an illustration of one channel of a phase vocoder or bandpass compressor.

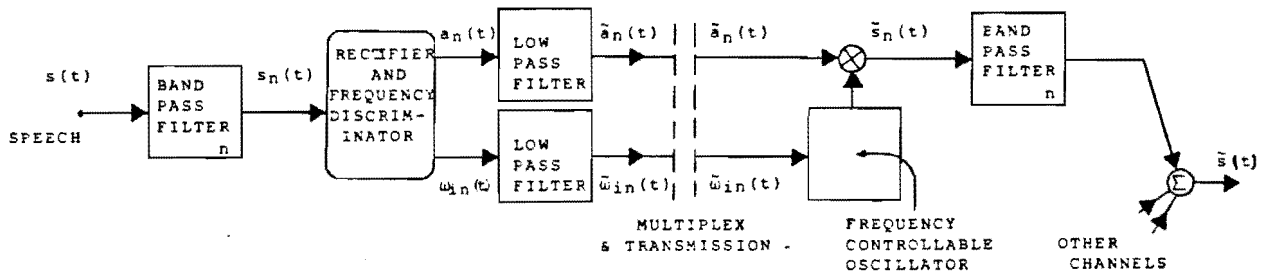


Fig. 1.14 Phase Vocoder Channel

Both systems analyse the output of each of the N bandpass filters in the comb to extract and transmit the instantaneous parameters of amplitude and frequency. If $S_n(t)$ is the output of the n^{th} filter ($n = 1,2,\ldots,N$), then

$$S_n(t) = a_n(t)\cos\phi_n(t) \qquad 1 \leq n \leq N \qquad \ldots (1.8)$$

where $a_n(t)$ and $\phi_n(t)$ are the instantaneous amplitude and phase of the n^{th} output. Instantaneous frequency is defined as the time derivative of instantaneous phase

$$\omega_i(t) = d(\phi(t))/dt \qquad \ldots (1.9)$$

and this calculation is performed for each $\phi_n(t)$.

It was known from the study of channel vocoders that $a_n(t)$ may be low pass filtered to 25 Hz and it was assumed that $\omega_{in}(t)$ could be treated similarly. Transmission of the 2N low pass signals is therefore achieved in approximately half of the original bandwidth.

At the receiver, each $\tilde{\omega}_{in}(t)$ can be integrated to form $\tilde{\phi}_n(t)$ or used directly to run a controllable oscillator. The output of the oscillator is amplitude modulated by $\tilde{a}_n(t)$ and the resulting signal passed through the n^{th} filter of a comb identical to that in the analyser. The synthesiser filter outputs are then summed to re-form the speech signal.

Since instantaneous frequencies appear as voltages, frequency division can be accomplished by simple attenuation. Dividing the instantaneous frequencies available at the analyser section of a phase vocoder has the effect of compressing the short time spectrum of the synthesised speech. In the time domain, the effect of such a frequency division is a slowing down (expansion) of the time waveform. Multiplication of the same instantaneous frequencies results in time compression or a speeding up of the time waveform.

The analyser section of a phase vocoder performs a type of short time Fourier analysis on the speech waveform. The channel outputs provide a continuous approximation to the speech amplitude and phase spectra, sampled in the frequency domain by the contiguous bandpass filters. Modern implementations of the phase vocoder operate on sampled speech data using Discrete Fourier Transform (DFT) or Fast Fourier Transform (FFT) techniques to generate the amplitude and phase spectra. In these cases, however, the instantaneous amplitude and frequency signals are necessarily discrete rather than continuous (Ref. 58, 59).

(1.5.4) SYSTEM PERFORMANCE - SPEECH QUALITY AND INTELLIGIBILITY

A major difficulty in weighing the performance of one speech transmission system against another is the inavailability of a precise quantative rating scale of speech quality. Speech quality estimates are formed by processing the responses of individuals to the "naturalness" of synthesised speech. Intelligibility scores are more precise but there are still the complicating factors of listener training and the context of the presented data (Ref. 60).

Synthesised speech can be graded on a rough speech quality scale consisting of four levels.

Broadcast quality is speech suitable for commercial radio commentary. It has a signal to noise ratio and harmonic distortion similar that of telephone quality speech, but a greater bandwidth. Depending on its ultimate use, the bandwidth may be 50 Hz to 10,000 Hz or 15,000 Hz.

Telephone quality is simply speech transmitted over a standard telephone bandwidth with a signal to noise ratio of greater than 30 dB and harmonic distortion less than 2 or 3%.

One step below telephone quality is communications quality. This is speech which may have lost some "naturalness" and may be audibly distorted, but is still very intelligible. Speaker recognition may also be difficult.

Synthetic quality speech typically sounds unnatural, possibly even mechanical. Talker recognition is very difficult and intelligibility may well be speaker and message dependent (Ref 49).

Some coding schemes are capable of synthetic quality transmission only. Among these are the channel vocoder, the formant vocoder and their variants. It must be noted however, that these vocoders may be working at bit rates as low as

500 bits per second. Other systems, such as the linear prediction vocoder, are on the borderline between synthetic and communications quality at bit rates around 5 kbits per second.

Analysis - synthesis telephony systems can achieve communications quality performance with the phase vocoder and voice excited vocoders working around 7 kbits per second. Waveform coding techniques provide communications quality at bit rates from 7 to 36 kbits per second. Low bit rate waveform coders are usually the most complex with sub-band types working around 10 kbits per second. At higher bit rates, simpler variations such as log companded P.C.M. will suffice for communications quality.

Both the phase vocoder and voice excited varieties are capable of telephone quality speech, working around 16 kbits per second. Telephone quality waveform coders work at higher bit rates and the trade-off between low rate and complexity applies as before.

Broadcast quality transmission usually requires waveform coding techniques working at rates above 64 kbits per second (Ref. 49).

The CCITT are currently considering speech coding schemes working at 32 kbits per second and 16 kbits per second. At present, delta modulation variations and A.D.P.C.M. seem likely as candidates for adoption as standards for 32 kbits per second coding. 16 kbits per second coding appears more difficult with coder complexity and processor time delay being important parameters (Ref 61, 62).

CHAPTER 2

(2.1) INTRODUCTION

The assumption common to all frequency division coding schemes is that the instantaneous frequency of a speech sub-band may be lowpass filtered without loss of significant information. To test this and other assumptions, it is necessary to examine the characteristics of typical instantaneous parameter waveforms.

(2.2) ANALYTIC SIGNALS

In order to assign physical meaning to the instantaneous parameters of an arbitrary bandpass signal, it is necessary to consider an "analytic" representation (or the "Pre-Envelope").

The real bandpass signal $s(t)$ is represented in terms of instantaneous amplitude $a(t)$ and phase $\phi(t)$ by equation

$$s(t)=a(t)\cos\phi(t). \qquad \ldots (2.1)$$

The corresponding analytic signal $\Psi(t)$ (or pre-envelope) is a complex valued function of time defined by

$$\Psi(t)=s(t)+j\hat{s}(t). \qquad \ldots (2.2)$$

For $\Psi(t)$ to be analytic, $s(t)$ and $\hat{s}(t)$ must be orthogonal and thus related by the Hilbert Transform pair

$$\left.\begin{aligned}\hat{s}(t)&= {}^{1}/_{\pi}\int_{-\infty}^{\infty} s(t)/(t-\tau)\ .d\tau\\ s(t)&={}^{-1}/_{\pi}\int_{-\infty}^{\infty} \hat{s}(t)/(t-\tau)\ .d\tau\end{aligned}\right\} \qquad \ldots (2.3)$$

(Ref. 65, 66, 67).

It can be seen from equation (2.3) that $\hat{s}(t)$ is the convolution of $s(t)$ with the function $1/\pi t$. In the frequency domain, this corresponds to phase shifting all positive Fourier frequency components by $-\pi/2$ radians and all negative Fourier frequency components by $+\pi/2$ radians. The real and imaginary parts of an analytic signal have the same power spectra and thus the same autocorrelation function.

An important property of the analytic signal can be established by examining the Fourier Transforms of s(t) and $\hat{s}(t)$.

$$s(t) \overset{FT}{\rightleftarrows} S(f) \text{ , for all frequencies } f \quad \ldots (2.4)$$

$$\hat{s}(t) \overset{FT}{\rightleftarrows} \begin{cases} -jS(f) & , f>0 \\ 0 & , f=0 \\ jS(f) & , f<0 \end{cases} \quad \ldots (2.5)$$

Combining equations (2.4) and (2.5), the Fourier Transform of $\Psi(t)$ is therefore

$$\Psi(f) = \begin{cases} 2S(f) & , f>0 \\ S(f) & , f=0 \\ 0 & , f<0 \end{cases} \quad \ldots (2.6)$$

which is one sided in the frequency domain.

Taking the original expression for the bandpass signal s(t), equation (2.1), and applying the Hilbert Transform relation, equation (2.3), gives the following orthogonal signal

$$\hat{s}(t)=a(t)\sin\phi(t). \quad \ldots (2.7)$$

Substituting equations (2.1) and (2.7) into equation (2.2)

$$\Psi(t)=a(t)\cos\phi(t)+ja(t)\sin\phi(t) \quad \ldots (2.8)$$

$$=a(t)e^{j\phi(t)} \quad \ldots (2.9)$$

From equation (2.9) it can be seen that the magnitude and phase of the analytic signal are the instantaneous amplitude and phase of the real waveform. The form of equation (2.9) suggests that the analytic signal is most effectively represented by a vector, magnitude a(t), rotating at the speed of the rate of change of phase, or instantaneous frequency $\omega_i(t)$

$$\omega_i(t)= {}^{d}/_{dt}(\phi(t)) \quad \ldots (2.10)$$

(Ref. 68, 69, 70, 71).

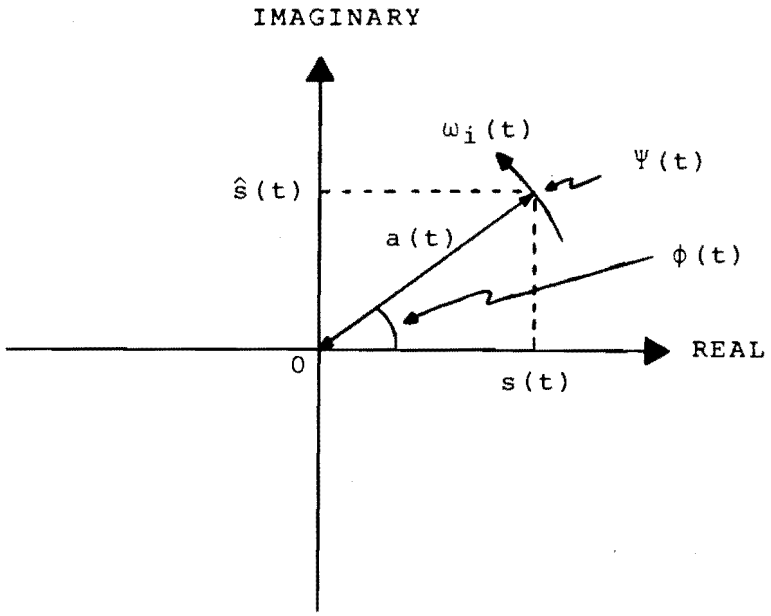


Fig. 2.1 Vector Representation of Analytic Signal

The analytic signal is fully defined by its real and imaginary components and access to both of these is required to calculate the instantaneous parameters of amplitude and phase.

(2.2.1) INSTANTANEOUS AMPLITUDE

From equations (2.1) and (2.7), it can be seen that, the square of the instantaneous amplitude function is

$$a^2(t)=s^2(t)+\hat{s}^2(t) \qquad \ldots (2.11)$$

Some idea of the bandwidth of this instantaneous waveform can be obtained by treating $a^2(t)$ as the square of the magnitude of the analytic signal

$$a^2(t)=|\Psi(t)|^2 \qquad \ldots (2.12)$$

Recognising that

$$|\Psi(t)|^2=\Psi(t).\Psi^*(t) \qquad \ldots (2.13)$$

where $\Psi^*(t)$ is the complex conjugate of $\Psi(t)$, allows us to write the Fourier Transform of $|\Psi(t)|^2$ as the convolution of $\Psi(f)$ and $\Psi^*(-f)$

$$\Psi(t) \overset{FT}{\rightleftarrows} \Psi(f) \qquad \ldots (2.14)$$

$$\Psi^*(t) \overset{FT}{\rightleftarrows} \Psi^*(-f) \qquad \ldots (2.15)$$

$$|\Psi(t)|^2 \overset{FT}{\rightleftarrows} \Psi(f) * \Psi^*(-f) \qquad \ldots (2.16).$$

As both $\Psi(f)$ and $\Psi^*(-f)$ are one sided in the frequency domain, the resulting spectrum will exhibit twice the bandwidth of the analytic signal. This is equivalent to the bandwidth of the real signal s(t). (Ref. 71)

In summary, the instantaneous amplitude waveform will be a positive function of time and the square of the waveform will be bandlimited to the bandwidth of the real signal s(t). After the square rooting operation, however, there is no reason to expect a(t) to be bandlimited.

(2.2.2) INSTANTANEOUS FREQUENCY

Instantaneous frequency is most clearly defined as the phase velocity of the analytic vector. In terms of the real and imaginary components, it is expressed

$$\omega_i(t) = {}^{d}/_{dt}\{\tan^{-1}(\hat{s}(t)/s(t))\} \qquad \ldots (2.17)$$

As it is possible to confuse definitions, it must be stressed that for a general signal s(t), the excursions of the time waveform $\omega_i(t)$ do not correspond to components of the short time Fourier amplitude spectrum. However, some correspondence may be detected if s(t) has very narrow bandwidth or for certain weighted averages of $\omega_i(t)$. (Ref. 72, 73)

There is no reason to expect $\omega_i(t)$ to be bandlimited and instantaneous frequency characteristics are probably

best demonstrated by examples or models.

(2.3) SPEECH MODELS

Although speech is difficult to represent mathematically, it can be approximated by a set of simple models.

(2.3.1) SINGLE SINUSOID

The most fundamental model is a sinusoid. Although it carries no information, it can be thought of as a single spectral component from the line spectrum of a voiced sound, or the output of one channel of a Phase Vocoder.

With no amplitude or angle modulation, s(t) is a constant level sinusoid of angular frequency ω radians/second.

$$s(t)=A\cos\omega t \qquad \ldots (2.18)$$

The analytic signal

$$\Psi(t)=Ae^{j\omega t} \qquad \ldots (2.19)$$

is a vector of length A rotating at a constant rate of ω radians/second.

$$\begin{aligned}\omega_i(t) &= {}^{d}/_{dt}(\omega t)\\ &= \omega \qquad \ldots (2.20)\end{aligned}$$

Figures 2.2.1 and 2.2.2 show the orthogonal waveforms s(t) and $\hat{s}(t)$. By the transform equation (2.3),

$$\hat{s}(t)=A\sin\omega t \qquad \ldots (2.21)$$

Both of the instantaneous parameters are represented by constants with respect to time (figures 2.2.3 and 2.2.4).

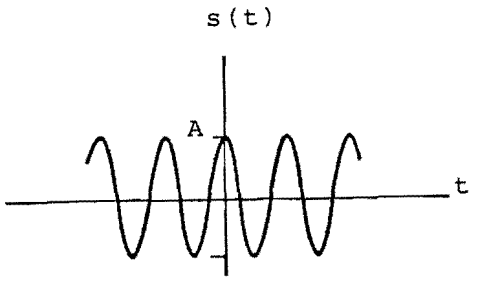


Fig. 2.2.1

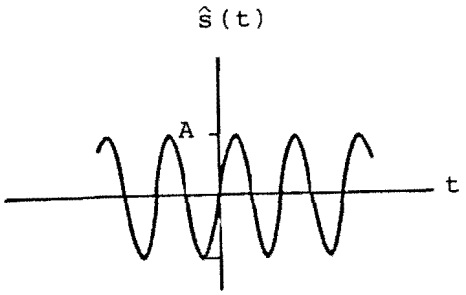


Fig. 2.2.2

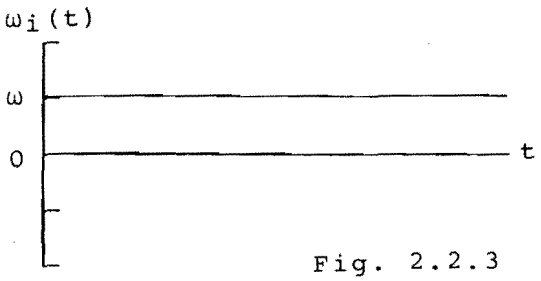


Fig. 2.2.3

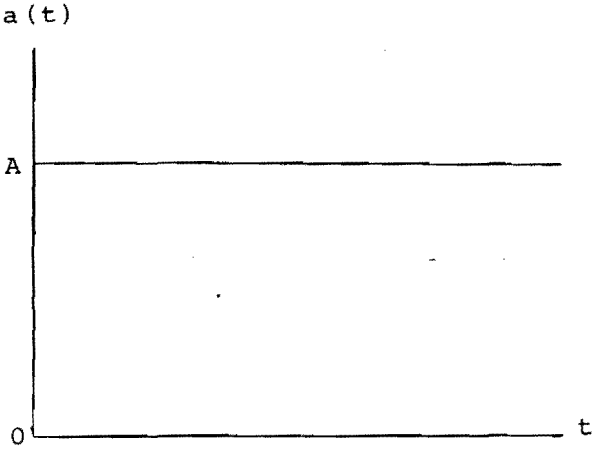


Fig.2.2.4

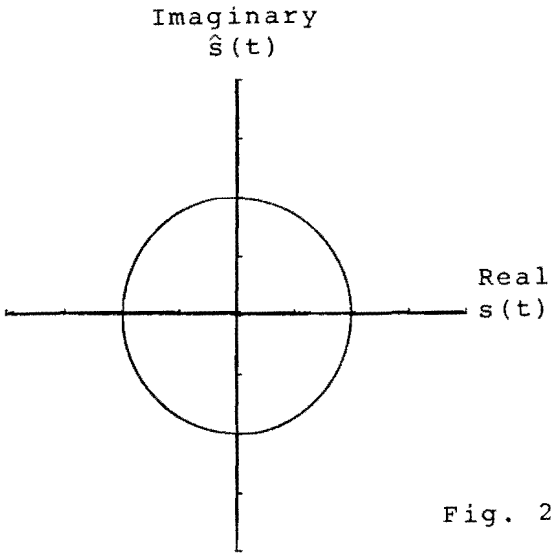


Fig. 2.2.5

Fig. 2.2 Analysis of a Cosine

In this case, Fourier and instantaneous frequencies correspond and frequency scaling is possible by simple attenuation or amplification of $\omega_i(t)$ before signal reconstruction.

Figure 2.2.5 is the locus of the tip of the vector $\Psi(t)$. The real signal s(t) can be plotted on the same set of axes by considering the relation

$$A\cos\omega t = {}^{A}/_{2}(e^{j\omega t}+e^{-j\omega t}) \qquad \ldots (2.22)$$

The pair of counter rotating vectors ${}^{A}/_{2}e^{j\omega t}$ and ${}^{A}/_{2}e^{-j\omega t}$, corresponding to positive and negative frequency spectral lines, add to form a resultant which never leaves the real axis. This is illustrated in figure 2.3.

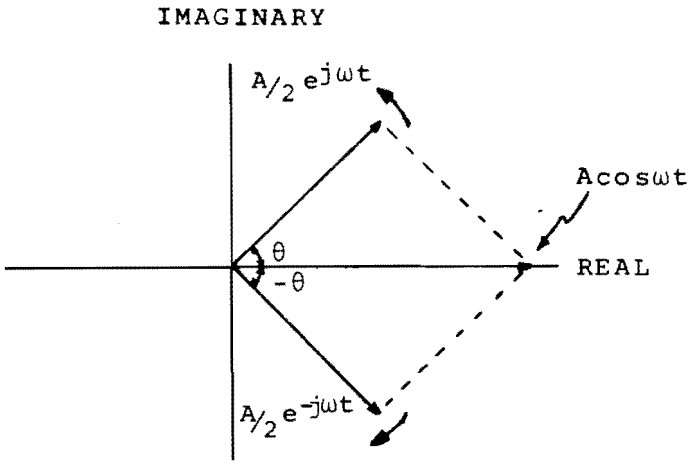


Fig. 2.3 Vector Representation of a Real Signal

(2.3.2) TWO SINUSOID SIGNAL

Any realistic channel of a phase vocoder or sub-band of a frequency division vocoder will pass more than one spectral line from a voiced spectrum. Unwieldy mathematical descriptions, however, will limit initial interest to the case of two unmodulated sinusoids with constant phase difference θ.

$$s(t)=A\cos\omega_1 t+B\cos(\omega_2 t+\theta) \qquad \ldots (2.23)$$

The corresponding formulae for instantaneous amplitude, phase and frequency are

$$a(t)=(A^2+B^2+2AB\cos(\omega_d t+\theta))^{\frac{1}{2}} \qquad \ldots (2.24)$$

$$\phi(t)=\omega_1 t+\tan^{-1}\left[\frac{(B/A)\sin(\omega_d t+\theta)}{1+(B/A)\cos(\omega_d t+\theta)}\right] \qquad \ldots (2.25)$$

$$\omega_i(t)=\omega_1+\frac{\omega_d(B/A)^2+\omega_d(B/A)\cos(\omega_d t+\theta)}{1+(B/A)^2+2(B/A)\cos(\omega_d t+\theta)} \qquad \ldots (2.26)$$

where ω_d is the difference frequency

$$\omega_d=\omega_2-\omega_1 \qquad \ldots (2.27)$$

Equations (2.24) to (2.26) are developed from (2.23) and the corresponding equation for $\hat{s}(t)$ in Appendix (A).

(2.3.2.1) INSTANTANEOUS AMPLITUDE

The instantaneous amplitude equation (2.24) of the two sinusoid signal is a positive function of time which exhibits one dip per cycle. It is periodic over the time T where

$$T=\frac{2\pi}{\omega_d}\ \text{seconds} \qquad \ldots (2.28)$$

The function can only reach zero when

$$A=B \qquad \ldots (2.29)$$

and $$(\omega_d t+\theta)=\pi \qquad \ldots (2.30)$$

Under these conditions, a(t) becomes a rectified cosine

$$a(t)=2A\left|\cos((\omega_d t+\theta)/2)\right| \qquad \ldots (2.31)$$

In the trivial case when B is zero, a(t) reverts to the constant envelope of a single sinusoid

$$a(t)=(A^2)^{\frac{1}{2}} \qquad \ldots (2.32)$$

(2.3.2.2) INSTANTANEOUS FREQUENCY

The instantaneous frequency waveform equation (2.26) is also periodic at the difference frequency ω_d. It is best described as a function of the two variables ρ and ξ where

$$\rho=(B/A) \qquad \ldots (2.33)$$

and $$\xi=(\omega_d t+\theta) \qquad \ldots (2.34)$$

Initially we will deal with the case of a two sinusoid signal with the major spectral component situated at ω_1 (equation (2.23)). ρ and ξ are therefore limited to the ranges $0 \leq \rho < 1$ and $0 \leq \xi \leq 2\pi$. Ignoring the constant additive term ω_1 and multiplying term ω_d in equation (2.26), figure 2.4 illustrates possible waveshapes of the function $f(\rho,\xi)$ where

$$f(\rho,\xi)=\frac{\rho^2+\rho\cos(\xi)}{1+\rho^2+2\rho\cos(\xi)} \qquad 0 \leq \rho \leq 1 \qquad \ldots (2.35)$$

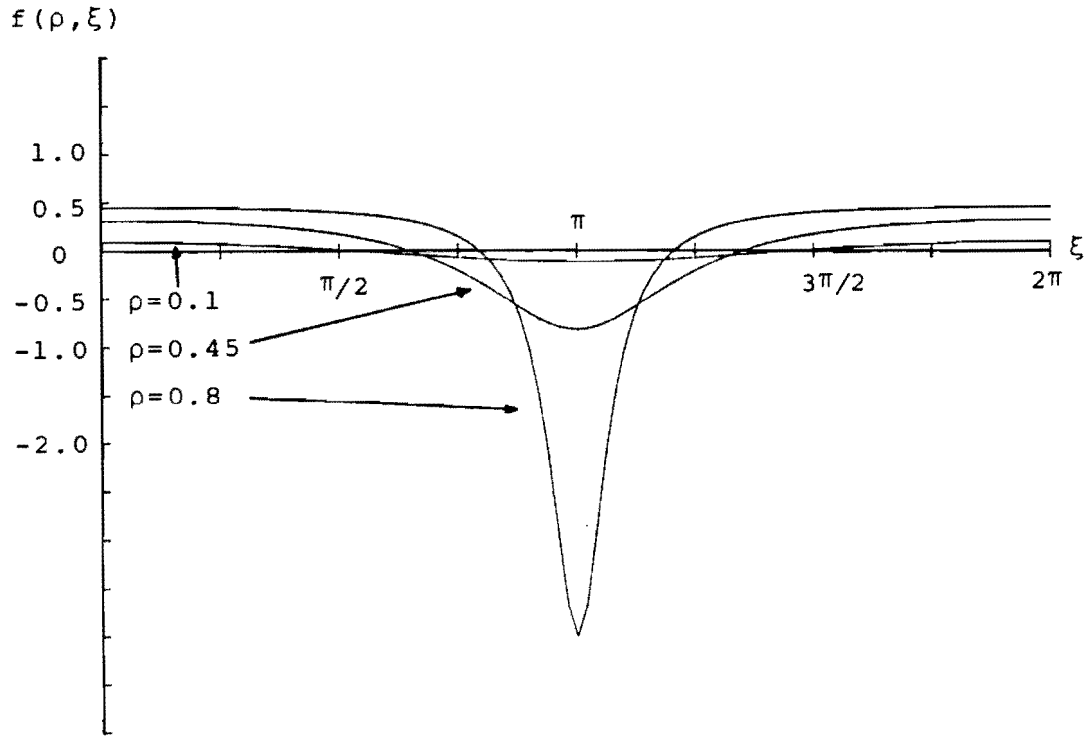


Fig. 2.4 Instantaneous Frequency Waveshapes

The basic shape is a dip at $\xi=\pi$. The depth and width of the dip are dependent on the variable ρ. As ρ tends to 1, the dip becomes deeper and narrower, in the limit becoming infinitely deep and of zero width.

The average value of $f(\rho,\xi)$ over one cycle of ξ is

$$\overline{f(\rho,\xi)} = \tfrac{1}{2\pi}\int_0^{2\pi} \frac{\rho^2+\rho\cos(\xi)}{1+\rho^2+2\rho\cos(\xi)} \,.d\xi \qquad \ldots (2.36)$$

$$= 0 \text{ , for } \rho<1 \quad \text{(Ref. 74)}$$

Expressing equation (2.26) for the instantaneous frequency of the two sinusoid model in terms of $f(\rho,\xi)$

$$\omega_i(t)=\omega_1+\omega_d.f(\rho,\xi) \qquad \ldots (2.37)$$

makes it clear that the average instantaneous frequency corresponds to the Fourier frequency of the stronger spectral component.

$$\overline{\omega_i(t)}=\omega_1 \qquad \ldots \quad (2.38)$$

When the spectral component at ω_2 is allowed to become dominant, ρ becomes restricted to the range $1 < \rho < \infty$. Figure 2.5 illustrates possible waveshapes of $f(\rho,\xi)$ when $\rho>1$.

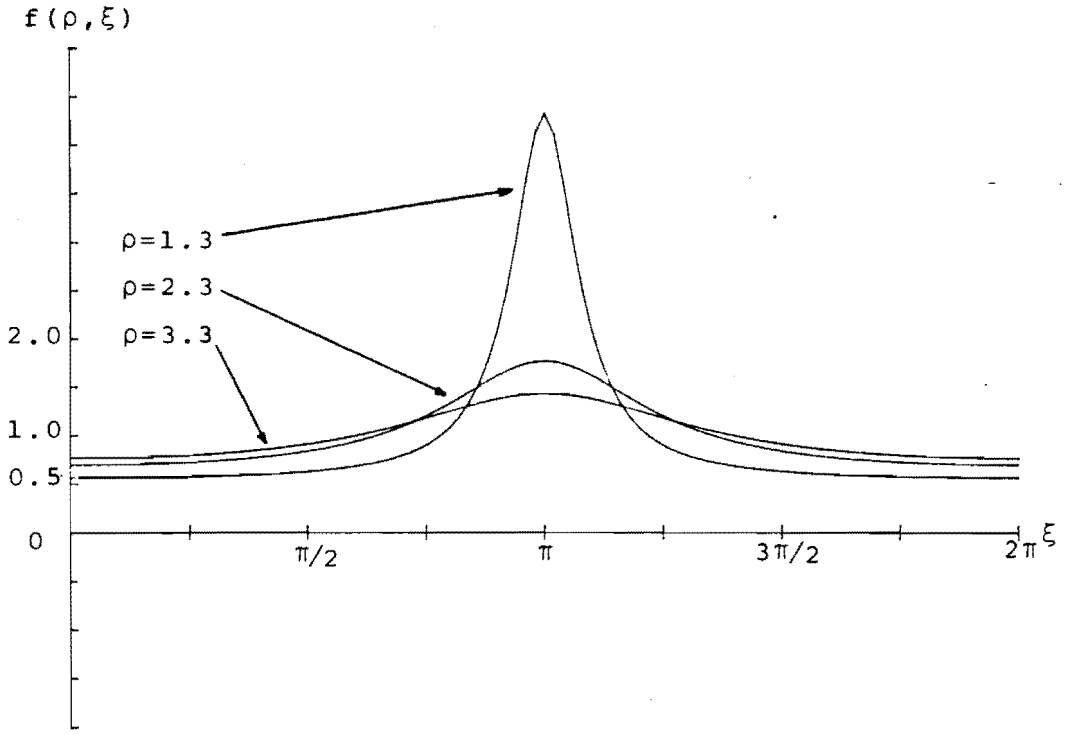


Fig. 2.5 Instantaneous Frequency Waveshapes

The family of curves is identical to that of figure 2.4 except that they are reflected in the line $f(\rho,\xi)=0.5$. As ρ tends to 1 (from above), $f(\rho,\xi)$ becomes a large narrow spike at $\xi=\pi$.

The average value of $f(\rho,\xi)$ over one cycle is easily deduced by comparing figures 2.4 and 2.5 and applying the solution of equation (2.36):

$$\overline{f(\rho,\xi)}=1 \quad , \quad \rho>1 \qquad \ldots \quad (2.39)$$

Substituting this result into equation (2.37) gives

$$\begin{aligned}\overline{\omega_i(t)} &= \omega_1+\omega_d \\ &= \omega_2 \qquad \ldots \quad (2.40)\end{aligned}$$

Once again the average value of instantaneous frequency corresponds to the Fourier frequency of the dominant spectral component.

Appendix (B) contains the results of a Fourier Series analysis of the instantaneous amplitude and frequency waveforms generated in the two sinusoid test case. Harmonic content, and thus effective bandwidths are dependent on ρ, with the amplitudes of higher harmonics increasing as ρ approaches 1 from above or below (Ref. 74).

(2.3.2.3) EXAMPLE - FUNDAMENTAL AND SECOND HARMONIC

The presence of the second spectral component in a two sinusoid signal distorts the analytic vector locus from the perfect circle of a pure sinusoid. The form of distortion can be illustrated by considering the case of a fixed amplitude fundamental component with variable second harmonic

$$s(t)=\cos\omega_1 t+B\cos 2\omega_1 t \qquad \ldots \quad (2.41)$$

Figure 2.6.1 shows one cycle of the instantaneous frequency of s(t) for $B<<1$. As expected, the addition of a small second harmonic component has caused a slight dip in the

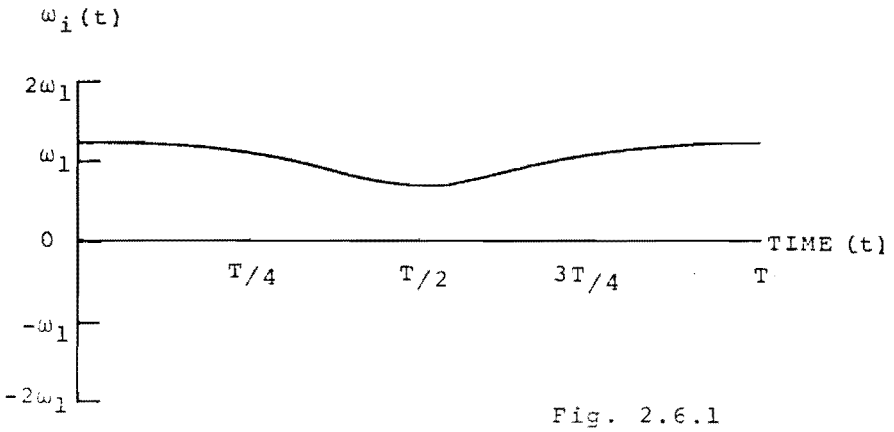


Fig. 2.6.1

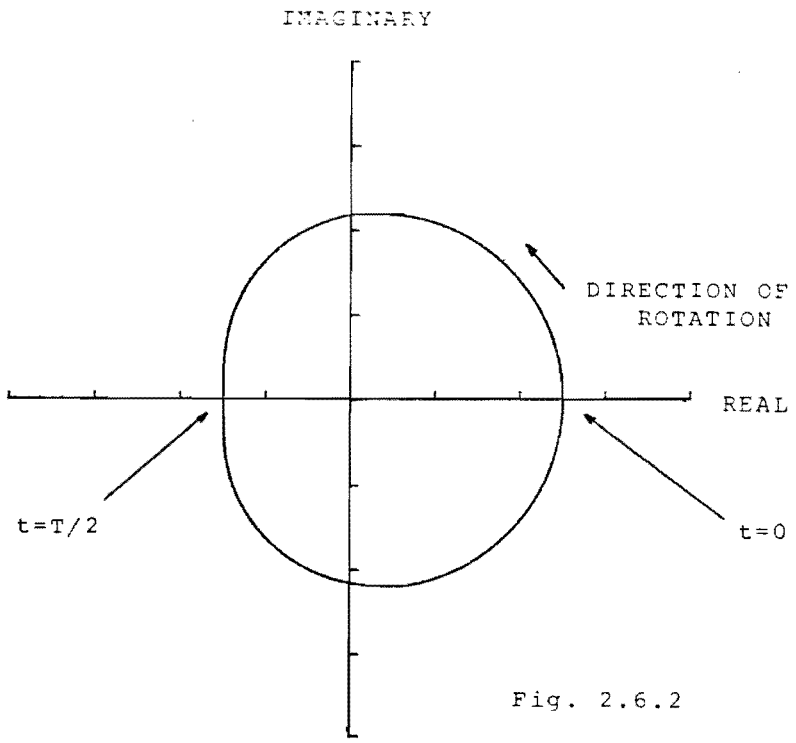


Fig. 2.6.2

Fig. 2.6 Curves for B << 1

waveform. The dip occurs at time T/2, where $T=2\pi/\omega_1$ seconds.

The corresponding dip in instantaneous amplitude is visible on the vector plot (figure 2.6.2) and it also occurs at time T/2. The vector completes a rotation in T seconds, but does not rotate at constant angular velocity (instantaneous frequency).

Increasing the magnitude of the second harmonic component ($B<1$) has the effect of deepening the instantaneous frequency trough. Considering equation (2.37), it will be seen that, for suitable values of ω_1 and ω_d, $\omega_i(t)$ can become negative. This is illustrated for one cycle of instantaneous frequency in figure 2.7.1.

The concept of negative instantaneous frequencies is most easily reconciled by treating them as reversals in the direction of vector rotation. Such a reversal is seen as an inner loop in the vector plot, figure 2.7.2. Once again, the instantaneous amplitude reaches a minimum at time T/2, at which point the vector is rotating clockwise.

Inner loops are therefore the result of negative instantaneous frequencies and simultaneous dips of instantaneous amplitude. The largest possible inner loop is one which just intersects the origin. If this occurs, then $a(t)=0$ once per cycle and from equation (2.24), $B=1$. The condition of both harmonics being of equal magnitude leads to $\omega_i(t)$ being undefined at time T/2.

If B is increased to the point where the second harmonic is dominant ($B>1$), the instantaneous frequency exhibits one positive spike per cycle (figure 2.8.1). As expected, the average instantaneous frequency is now $2\omega_1$. The vector plot (figure 2.8.2) provides a physical explanation for the quantum step of average instantaneous frequency from ω_1 to $2\omega_1$.

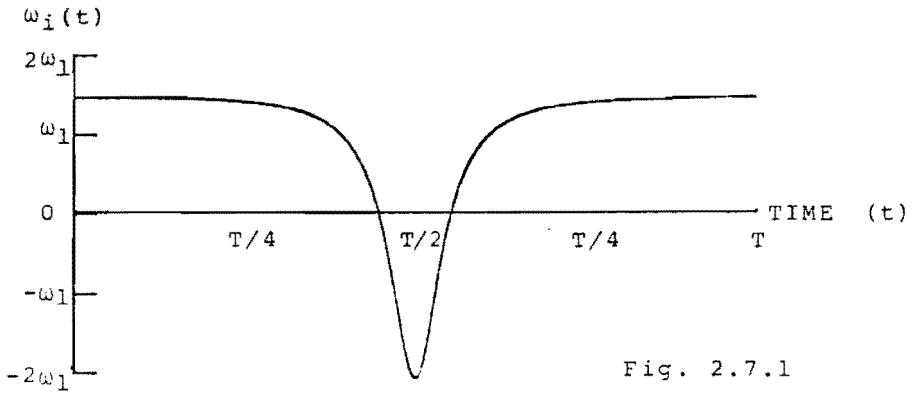


Fig. 2.7.1

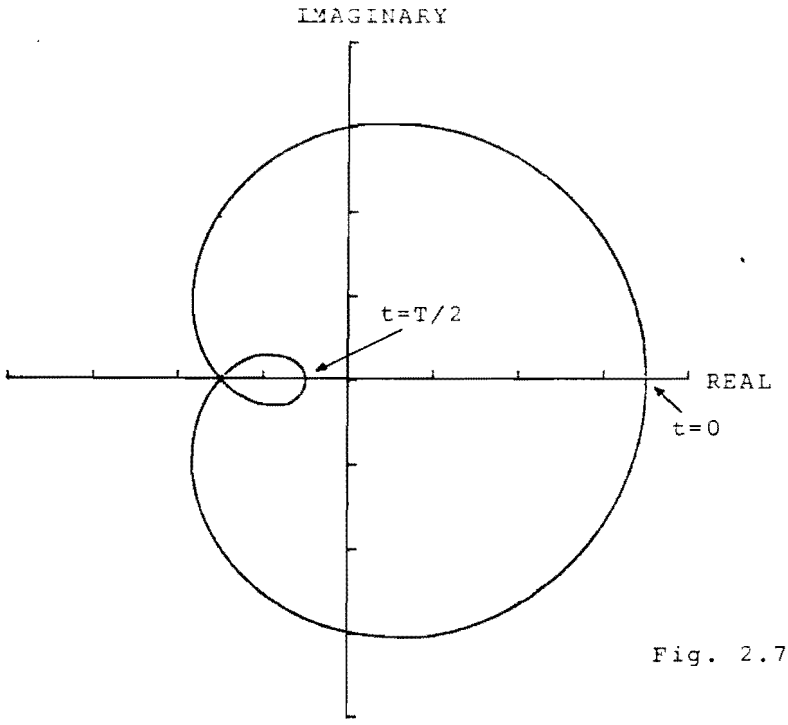


Fig. 2.7.2

Fig. 2.7 Curves for B < 1

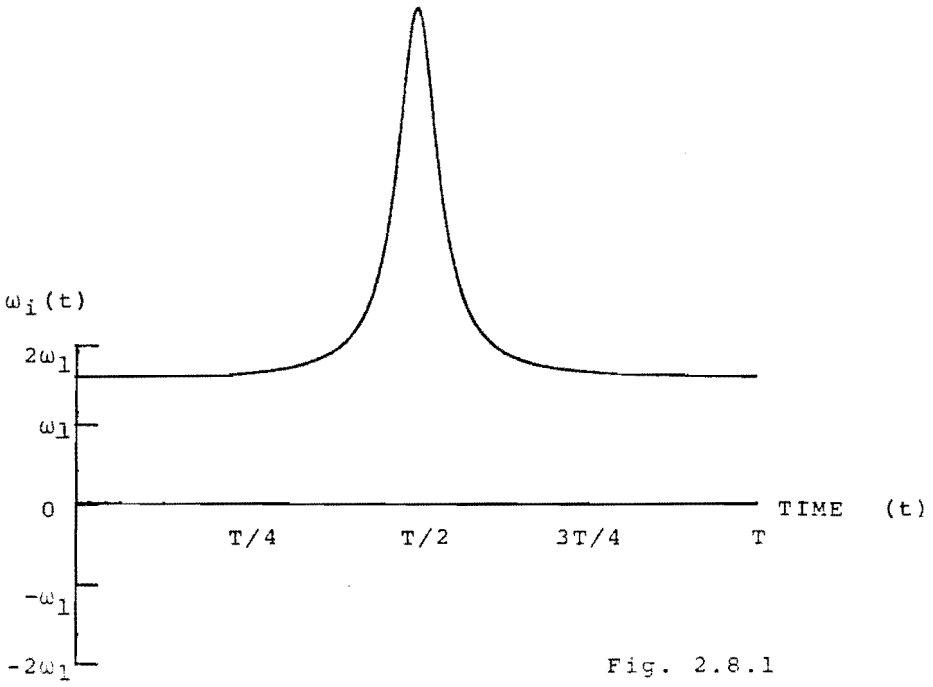


Fig. 2.8.1

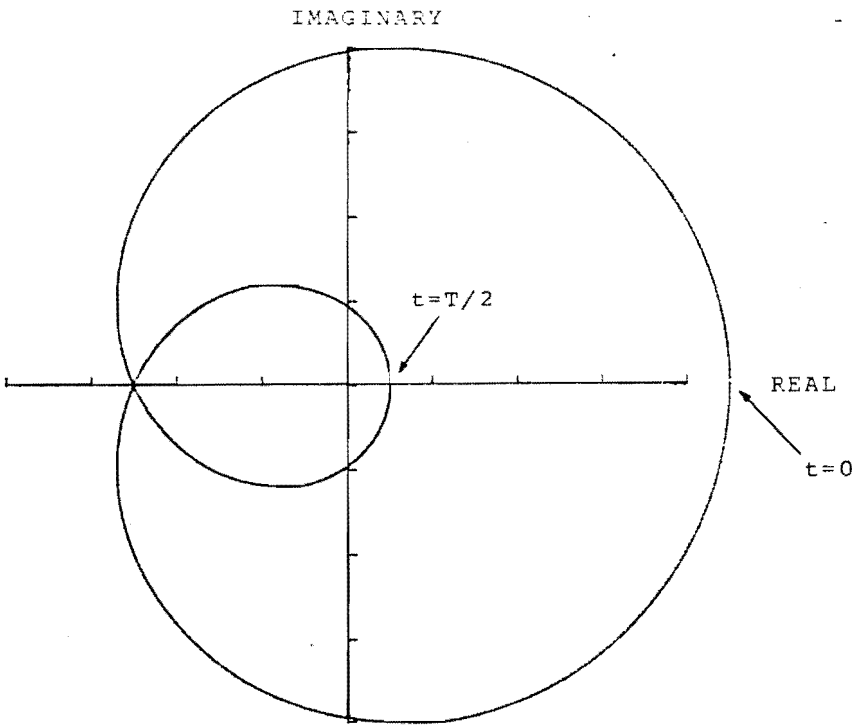


Fig. 2.8.2

Fig. 2.8 Curves for B > 1

The inner loop of figure 2.7.2 has now expanded to encompass the origin. At time T/2, the instantaneous amplitude dip does not reach zero, but the vector has completed one full encirclement of the origin. Over one cycle of instantaneous frequency, the vector loops the origin twice, resulting in an average angular velocity of 4π radians in T seconds or $2\omega_1$ radians per second. Although $\omega_1(t)$ is undefined at time T/2 for B=1, the change of average instantaneous frequency from ω_1 to $2\omega_1$ occurs immediately B becomes greater than 1.

If B is allowed to become much greater than 1, the instantaneous frequency curve will be less spiked and more constant around $2\omega_1$. The vector plot will become circular, but the vector will rotate twice in T seconds.

(2.3.2.4) SUMMARY

Several important factors arise from the two sinusoid analysis. The instantaneous frequency of a very simple bandlimited signal is unbandlimited and possesses infinite dynamic range. The average value corresponds to the frequency of the dominant spectral component and spike or trough fluctuations occur at the rate of the difference frequency between components. Large positive or negative instantaneous frequency excursions correspond to low instantaneous amplitude values.

Analysis of vector plots explains the significance of negative instantaneous frequencies and physically demonstrates the sudden change of average instantaneous frequency when an inner loop expands to encompass the origin. This effect is analogous to stronger signal capture in frequency modulation transmission theory. (Ref. 75).

Although the instantaneous frequency signal is essentially unbandlimited, the output of a real phase vocoder channel would probably be quite well behaved. Any large peaks and troughs of $\omega_1(t)$ would be attenuated by the severe low pass filter and the output will be an average instantaneous frequency corresponding approximately to the frequency of the dominant component.

Low pass filtering of instantaneous amplitude will remove any sharp dips, leaving a signal which is proportional to the average energy in the passband.

(2.3.3) BANDLIMITED NOISE

When the vocal utterance is a stationary unvoiced sound, the signal in a sub-band channel may be approximated by narrow-band Gaussian noise n(t). Instantaneous amplitude, phase and frequency of such signals are best described by their probability density functions.

The instantaneous amplitude or envelope of narrow-band Gaussian noise fits the well known Rayleigh distribution. If a_t is the value of instantaneous amplitude at time t and σ the standard deviation of the Gaussian random variable n(t), then the probability density function is

$$p(a_t) = \frac{a_t}{\sigma^2} \exp\left(\frac{-a_t^2}{2\sigma^2}\right), a_t \geq 0 \qquad \ldots (2.42)$$

(Ref. 67)

Equation (2.42) is plotted in figure 2.9

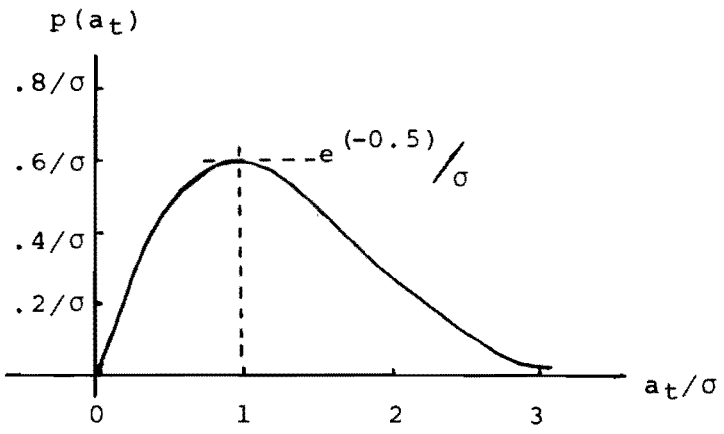


Fig. 2.9 Rayleigh Distribution

The peak value occurs at $a_t=\sigma$ and, since instantaneous amplitude is a positive function, the distribution is zero for negative values of a_t.

Instantaneous phase of narrow-band Gaussian noise is uniformly distributed in the range $0 \leq \phi(t) \leq 2\pi$.

$$p(\phi_t) = 1/2\pi \ , \ 0 \leq \phi_t \leq 2\pi \qquad \ldots \ (2.43)$$

Instantaneous frequency, however, conforms to a more complex distribution. It is dependent on the centre frequency of the channel filter ω_m and the channel bandwidth $\Delta\omega$. If ω_t is the value of instantaneous frequency at time t then the probability density function is

$$p(\omega_t) = \frac{\sigma^2}{2} \left\{ (\omega_t-\omega_m)^2+\sigma^2 \right\}^{-3/2} \qquad \ldots \ (2.44)$$

(Ref. 76,77)

The symbol σ^2 in equation (2.44) refers to the spectral variance of the signal n(t). This can be defined

$$\sigma^2 = \pi^2 \lambda_o^2 - \omega_m^2 \qquad \ldots \ (2.45)$$

where λ_o is the rate at which n(t) crosses the zero axis. From the work of Rice (Ref. 78) it is known that if $N(\omega)$ is the spectral density of the noise, then

$$\lambda_o = \frac{1}{\pi}\left(\frac{\int_{-\infty}^{\infty} \omega^2 N(\omega).d\omega}{\int_{-\infty}^{\infty} N(\omega).d\omega}\right)^{\frac{1}{2}} \quad \ldots (2.46),$$

and equation (2.45) can be rewritten

$$\sigma^2 = \frac{\int_{-\infty}^{\infty} \omega^2 N(\omega).d\omega}{\int_{-\infty}^{\infty} N(\omega).d\omega} - \omega_m^2 \quad \ldots (2.47)$$

Equation (2.47) is the definition of the bandwidth of a narrow-band Gaussian process and this relationship allows us to redefine spectral variance in terms of "equivalent rectangular" bandwidth

$$\sigma^2 = \frac{\Delta\omega^2}{12} \quad \ldots (2.48)$$

(Ref. 79).

Rewriting equation (2.44) in terms of "equivalent rectangular" bandwidth gives

$$p(\omega_t) = \frac{\Delta\omega^2}{24}\left\{(\omega_t-\omega_m)^2 + \frac{\Delta\omega^2}{12}\right\}^{-3/2} \quad \ldots (2.49)$$

$(\omega_m \geq \Delta\omega/2)$

The characteristics of the probability density function of instantaneous frequency are therefore fully described in terms of the noise bandwidth and centre frequency. The shape of the function is illustrated in figure 2.10.

The curve is symmetrical about ω_m, the peak value being $\frac{1}{2}(\sqrt{12}/\Delta\omega)$. The probability of negative instantaneous frequencies may be very small but is always greater than zero when $\Delta\omega > 0$

$$\text{Prob}\{\omega_t<0\} = \tfrac{1}{2}\left\{1 - \frac{\omega_m}{\left\{(\Delta\omega^2/12)+\omega_m^2\right\}^{\frac{1}{2}}}\right\} \quad \ldots (2.50)$$

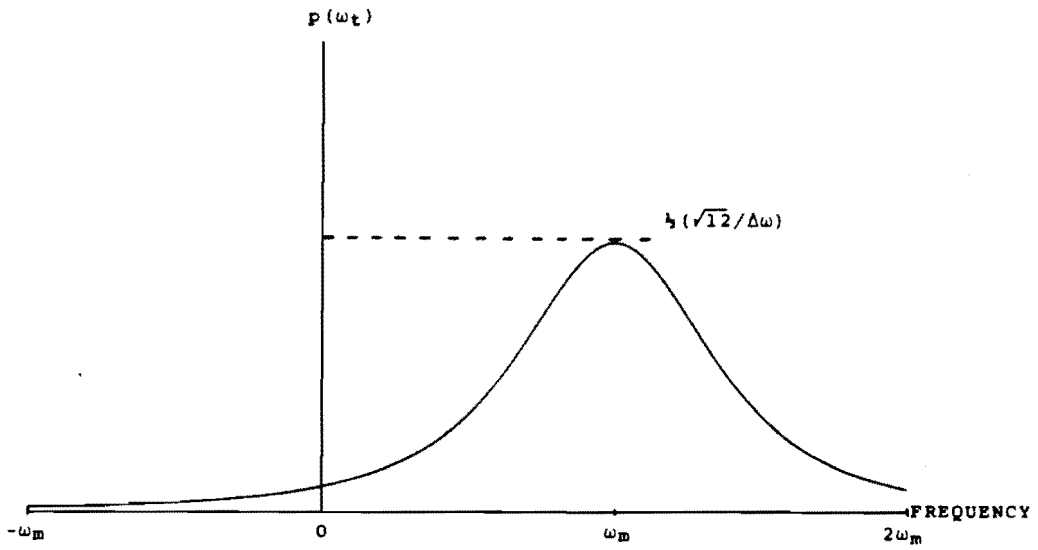


Fig. 2.10 Instantaneous Frequency Probability Density Function

Low pass filtering time waveforms corresponding to instantaneous amplitude and frequency distributions will effectively remove extreme values and in the case of $\omega_i(t)$, will compress the distribution around its centre value ω_m. If the bandpass channel filters of a phase vocoder are of sufficiently narrow bandwidth, such low pass filtering would not be expected to cause severe degradation of the reconstructed signal.

(2.4) COMPLEX DEMODULATION

In order to extend a study of instantaneous parameters beyond the limitations of mathematically well behaved models, it is necessary to study the characteristics of real signals. This requires definition of a complex demodulation scheme, ideally capable of handling a full telephone bandwidth speech signal.

(2.4.1) AMPLITUDE DEMODULATION

Extraction of the instantaneous amplitude waveform

from a real signal is basically a matter of accurate envelope detection. This can be performed at baseband frequencies making use of a wideband Hilbert transformer and the relationship

$$a(t)=(s^2(t)+\hat{s}^2(t))^{\frac{1}{2}} \qquad \ldots (2.51)$$

It can also be performed on a frequency translated Single Sideband (SSB) version of s(t). The same mathematical relationship holds, but Hilbert transformation at high frequency does not require the complex transversal filtering techniques necessary at baseband.

(2.4.2) FREQUENCY DEMODULATION

Instantaneous frequency, or the speed of rotation of the analytic vector, is related to the real signal and its Hilbert transform by the equations

$$\omega_i(t)= {}^{d}/_{dt}\{\tan^{-1}(\frac{\hat{s}(t)}{s(t)})\} \qquad \ldots (2.52)$$

or

$$\omega_i(t)= \frac{s(t).\hat{s}'(t)-s'(t).\hat{s}(t)}{s^2(t)+\hat{s}^2(t)} \qquad \ldots (2.53)$$

where the prime superscript "'" in equation (2.53) indicates differentiation with respect to time.

Frequency demodulation by implementation of either equation (2.52) or (2.53) is significantly more complicated than the discrimination method employed in standard FM receivers. The first step in a simple demodulation scheme is to hard limit (clip) the received signal. This effectively removes all amplitude variation and any remaining information can be thought of purely in terms of spacings between zero crossings or variations of the zero crossing rate. If the modulating signal, $\omega_i(t)$, is well controlled, it may be fully defined in terms of the zero crossings of the frequency modulated signal.

An estimate of instantaneous frequency based on zero crossing positions has the form

$$\text{Estimated } \omega_i(t) = \left| d/dt\left\{\tan^{-1}\left(\frac{\hat{s}(t)}{s(t)}\right)\right\}\right| \quad \text{(Ref. 80)} \quad \ldots (2.54)$$

This result implies that instantaneous frequency can be recovered from knowledge of a real signals zero crossings provided the term within the modulus signs (equation 2.54) is never negative. Such a demodulation scheme is suitable for broadcast FM signals, but for a general baseband signal, the estimate is erroneous.

The error can be illustrated by examining the average instantaneous frequency of a two sinusoid signal

$$s(t)=A\cos\omega_1 t+B\cos\omega_2 t \quad \ldots (2.55)$$

$\omega_i(t)$ begins its negative frequency excursions when the ratio B/A becomes just greater than ω_1/ω_2 and exhibits negative values over the range

$$\omega_1/\omega_2 < B/A < 1.$$

Figure 2.11 is a plot of average instantaneous frequency versus amplitude ratio B/A calculated from zero crossing information from the signal

$$s(t)=A\cos(2400\pi t)+B\cos(3200\pi t) \quad \ldots (2.56)$$

Previous analysis has shown that the average instantaneous frequency steps between 1.2kHz and 1.6kHz at the point B/A=0dB. One explanation for the high readings in the range

$$-2.5\text{dB} \leq B/A \leq 0\text{dB}$$

is rectification of negative instantaneous frequency excursions in the estimator equation (2.54).

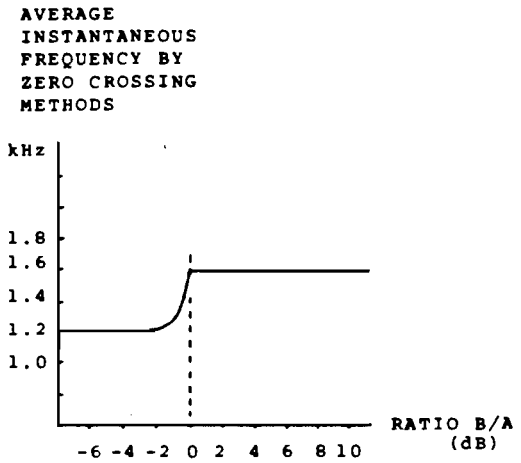


Fig. 2.11 $\overline{\omega_i(t)}$ by Zero Crossing Methods

This effect is illustrated in figure 2.12 (Ref. 81, 82).

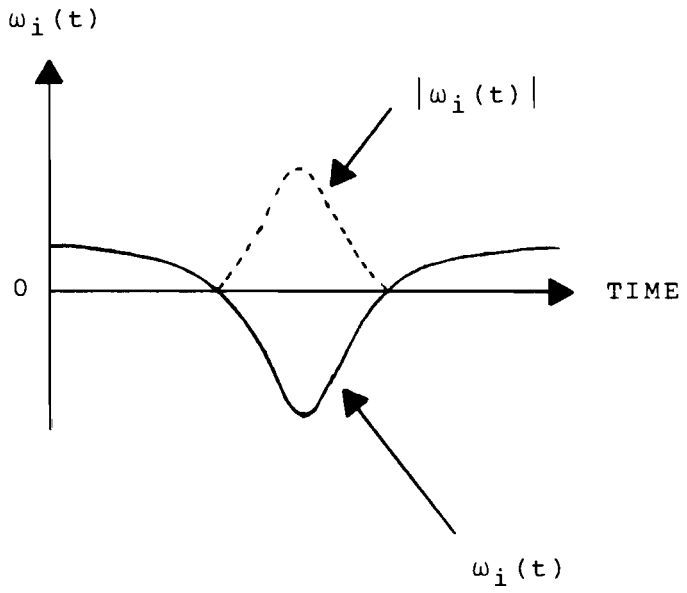


Fig. 2.12 Rectification of $\omega_i(t)$

Although it is obvious that successful complex demodulation of a baseband signal will require an implementation of either equation (2.52) or (2.53), it appears that the zero crossings of a signal carry significant instantaneous frequency information. This phenomenon is related to the fact that hard limited (infinite peak clipped) baseband speech is highly intelligible (Ref.83). Such observations have important consequences and have lead to an alternative mathematical interpretation of instantaneous waveforms.

(2.5) ZEROS OF REAL AND ANALYTIC SIGNALS

It is clear from equation (2.54) that, in general, zero crossings provide only a partial description of a band-limited signal. In order to represent the amplitude information stripped by the clipping operation, it is necessary to consider a signals real zeros (zero crossings) and its complex zeros.

The complex zeros of a real waveform are not as physically obvious as zero crossings, but usually correspond to visible waveform dips and fluctuations. Although this makes them difficult to locate, they can be calculated from mathematical representations of a waveform by allowing the time variable to become complex

$$z = \tau + j\sigma \qquad \ldots (2.57)$$

In order to impart physical meaning to the concept of complex zeros, it is convenient to consider the small section of the real waveform e(t) illustrated in figure 2.13.1.

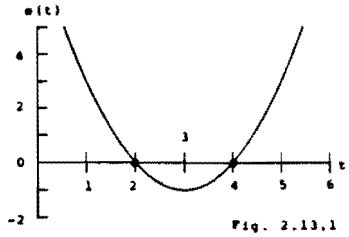


Fig. 2.13.1

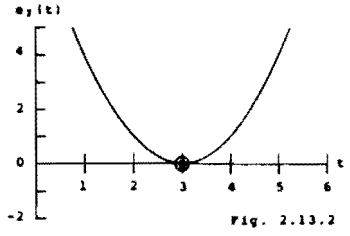


Fig. 2.13.2

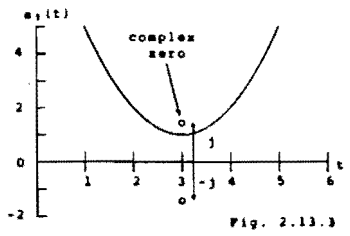


Fig. 2.13.3

Fig. 2.13 Waveform Segments

This particular waveform can be approximated over the time interval of interest by the second order algebraic polynomial

$$e_1(t)=t^2-6t+8 \quad , \; 0 \leq t \leq 6 \qquad \ldots \; (2.58)$$

which exhibits zero crossings at t=2 and t=4. This is a pair of real zeros.

Adding a positive DC component (of appropriate magnitude) lifts the waveform to the point where the curve just touches the time axis at its minimum. The modified waveform segment is illustrated in figure 2.13.2 and is now approximated by

$$e_2(t)=t^2-6t+9 \quad , \; 0 \leq t \leq 6 \qquad \ldots \; (2.59)$$

The resulting zero configuration is a second order zero at t=3.

Addition of a second DC component lifts the curve above the time axis, figure 2.13.3, and precludes zero crossings in the interval $0 \leq t \leq 6$. The waveform is now approximated by

$$e_3(t)=t^2-6t+10 \quad , \; 0 \leq t \leq 6 \qquad \ldots (2.60)$$

which exhibits a conjugate pair of complex zeros at $z = 3 \pm j$.

It is notable that each waveform segment displays two zeros over the range $0 \leq t \leq 6$ and that this feature is a function of waveshape rather than DC level.

In the case of $e_3(t)$, the segment corresponds to a waveform dip or fluctuation, the proximity of the minimum to the time axis being indicated by the magnitude of the imaginary component, $|\sigma|$, of the associated pair of complex zeros. In general, large dips produce complex zeros which are close to the real time axis and thus have small values of σ.

In his definitive papers (Ref. 84, 85) Voelker develops the theory of real and complex zeros to the point where the instantaneous amplitude and frequency of a waveform can be defined solely in terms of the zero positions of the associated analytic signal. The relations are

$$d/dt\{\ln(a(t))\}=d/dt\{\ln(a(0))\}+\sum \frac{r_n\tau_n}{\tau_n^2+\sigma_n^2} +\sum \frac{r_n(t-\tau_n)}{(t-\tau_n)^2+\sigma_n^2} \qquad \ldots (2.61)$$

and

$$\omega_i(t)=\omega_i(0)-\sum \frac{r_n\sigma_n}{\tau_n^2+\sigma_n^2}+\sum \frac{r_n\sigma_n}{(t-\tau_n)^2+\sigma_n^2} \qquad \ldots (2.62)$$

where the summations are over the total number of zeros and r_n is the order of the nth zero $z_n=\tau_n+j\sigma_n$.

The major consequence of equations (2.61) and (2.62) is that the analytic signal, and thus its real bandlimited projection, is fully described by zero positions. This leads directly to the conclusion that zeros are "informational

attributes" of a waveform upon which we can impose a set of rules similar to the Nyquist Sampling Theorem.

The zeros of a real, bandlimited signal are either pairs of zero crossings or complex conjugate pairs, but they must occur at an average rate equivalent to the Nyquist sampling frequency. This is normally defined to be twice the frequency of the highest frequency signal component.

Modelling the analytic signal as a vector spiralling along the time axis makes it apparent that zero crossings do not occur and subsequently that all zeros of an analytic signal are complex.

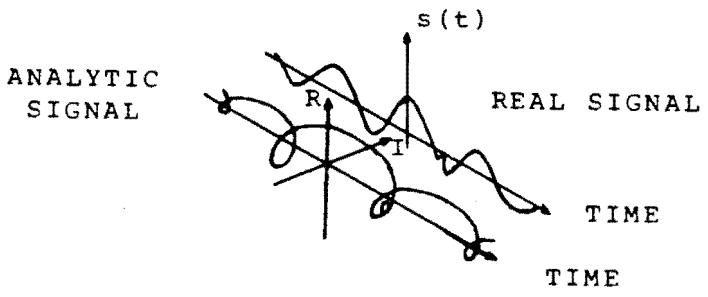


Fig. 2.14 Analytic Signal Model

The one sided spectral property (half bandwidth) of the analytic signal suggests that the zeros of an analytic signal occur at half the rate of those of the real projection. Such a reduction of zero occurrence without a loss of information can only be accomplished if one analytic signal zero is generated by each pair of real signal zeros. The zeros of an analytic signal, therefore, do not occur in conjugate pairs, but are single complex zeros in either the upper half of the complex plane (UHP, σ positive) or in the lower half of the complex plane (LHP, σ negative) (Ref. 84).

With this knowledge of the analytic signal zeros, it is possible to interpret the waveshapes predicted by equations (2.61) and (2.62).

The variable term in equation (2.61) for the time differentiated natural logarithm of instantaneous amplitude is

$$u(t) = \sum \frac{r_n \tau_n}{\tau_n^2 + \sigma_n^2} + \sum \frac{r_n (t-\tau_n)}{(t-\tau_n)^2 + \sigma_n^2} \qquad \ldots (2.63)$$

Taking into account the effects of integration and $\ln^{-1}$ operations on u(t), the presence of a complex zero will be indicated by a dip of the positive instantaneous amplitude function at time τ_n. As σ_n appears only as a squared term, the variation will be unaffected by whether the zero is UHP or LHP.

The variable term from equation (2.62) is

$$v(t) = \sum \frac{-r_n \sigma_n}{\tau_n^2 + \sigma_n^2} + \sum \frac{r_n \sigma_n}{(t-\tau_n)^2 + \sigma_n^2} \qquad \ldots (2.64)$$

The type and magnitude of instantaneous frequency variation is dependent on the sign and magnitude of σ_n. If σ_n is positive (UHP zero), v(t) is a "rise" or "spike". As $|\sigma_n|$ is reduced, the spike becomes higher and narrower, but maintains its area as

$$\int_{-\infty}^{\infty} \frac{\sigma_n}{(t-\tau_n)^2 + \sigma_n^2} \, .dt = \pi \qquad \ldots (2.65)$$

If σ_n is negative (LHP zero), the variation is in the form of a constant area dip which becomes steeper and narrower as $|\sigma_n|$ reduces.

At the time t=0, v(t)=0 and instantaneous frequency is determined by the constant $\omega_i(0)$.

The predicted dips of instantaneous amplitude and the constant area spike and dip variations of $\omega_i(t)$ correspond to those exhibited by instantaneous waveforms in the previous examples. However, it is now possible to say that these variations correspond in real time to analytic signal complex zero positions, with instantaneous frequency spikes indicating UHP zeros and dips indicating LHP zeros. The shape of instantaneous frequency spikes or dips gives an indication of $|\sigma_n|$.

Equations (2.61) and (2.62) are elegant relationships and are an exact means of obtaining instantaneous amplitude and frequency. Unfortunately, the requirement that we locate all of the analytic signal zeros makes them difficult to apply.

The next step of this analysis is, by means of an example, to relate analytic signal zeros to the zeros of its real projection and thus develop rules for analysis and prediction of instantaneous waveforms.

(2.5.1) EXAMPLE

Once again a two sinusoid test signal is employed, but for ease of calculation, this is the fundamental and third harmonic

$$s(t)=\cos\omega t-B.\cos 3\omega t \qquad \ldots (2.66)$$

Allowing time to become complex, the analytic signal corresponding to s(t) may be written

$$\begin{aligned}\Psi(z)&=e^{j\omega z}-B.e^{3j\omega z}\\&=e^{j\omega z}\{1-B.e^{2j\omega z}\}\end{aligned} \qquad \ldots (2.67)$$

The roots of the equation $\Psi(z)=0$ are therefore $e^{j\omega z}=0,\pm 1/\sqrt{B}$. Taking natural logarithms locates these roots in the complex time plane, z, and $e^{j\omega z}=\pm 1/\sqrt{B}$ maps to $z=-j\ln|1/\sqrt{B}|/\omega$ or $z=\pi/\omega-j\ln|1/\sqrt{B}|/\omega$. These roots repeat at intervals of $2n\pi/\omega$ along the real time axis.

The root $e^{j\omega z}=0$ leaves τ undefined and places σ at $+\infty$. Voelkers equations (2.61) and (2.62) predict that this analytic signal zero will cause no fluctuations of instantaneous amplitude or frequency, but will contribute a constant 2π radians per cycle to the value of $\omega_i(t)$. In this respect, it may be treated as a frequency shift of $+\omega$ radians per second. For the purposes of this analysis, such analytic signal zeros at $\sigma=+\infty$ will be termed "removed".

The zeros with imaginary component $\sigma=-\ln|1/\sqrt{B}|/\omega$ are LHP when $B<1$ and UHP when $B>1$.

Representing cosine terms in exponential form and permitting time to become complex allows the equation of the real signal (2.66) to be rewritten

$$s(z)=\tfrac{1}{2}(-B.e^{3j\omega z}+e^{j\omega z}+e^{-j\omega z}-B.e^{-3j\omega z}) \quad . . . \quad (2.68)$$

The roots of $s(z)=0$ are therefore found by factorising

$$-B.e^{6j\omega z}+e^{4j\omega z}+e^{2j\omega z}-B=0 \quad . . . \quad (2.69)$$

which yields

$$(e^{2j\omega z}+1)(-B.e^{4j\omega z}+(1+B).e^{2j\omega z}-B)=0 \quad . . . \quad (2.70)$$

The factor $(e^{2j\omega z}+1)$ locates one pair of zero crossings per cycle of $s(z)$ at $\tau=\pi/2\omega$ and $\tau=3\pi/2\omega$. These zero crossing positions are independent of the value of B.

Treating the second factor as a quadratic in $x=e^{2j\omega z}$ leads to the following solution for the remaining four zeros per cycle of s(z)

$$e^{j\omega z}=\pm\left(\frac{(1+B)\pm\sqrt{(-3B-1)(B-1)}}{2B}\right)^{\frac{1}{2}} \qquad \ldots (2.71)$$

When $B<1$, these zeros are complex conjugate pairs at $\tau=0$ and $\tau=\pi/\omega$. For $B>1$, they become four distinct zero crossings. The real and analytic signal zero positions are calculated for several values of B in Appendix (C).

In the trivial case when B=0, s(t) is a simple sinusoid. The single analytic signal zero per cycle is "removed" ($\sigma=+\infty$) and instantaneous amplitude and frequency are constants with the values $a(t)=1$, $\omega_i(t)=\omega$.

Substituting the value B=0 into equation (2.70) yields the expected pair of real signal zero crossings at $\tau=0$ and $\tau=\pi/\omega$. The average value of instantaneous frequency and the average rate of real signal zero crossing occurrence correspond in this case.

The vector plot corresponding to the analytic signal is a circle of radius 1.

The addition of a small third harmonic component, B=0.25, introduces periodic fluctuations of the instantaneous parameters at a rate of two per cycle of s(t), figure 2.15 This rate is proportional to the signal bandwidth, 2ω.

The instantaneous waveforms, figures 2.15.3 and 2.15.4, exhibit simultaneous dips which correspond in time with visible dips of the real waveform, figure 2.15.1. Plotting the locations of "non-removed" analytic signal complex zeros in figure 2.15.1 ($\sigma_c=-0.69/\omega$ for B=0.25) illustrates the relationship between these zero positions and real and

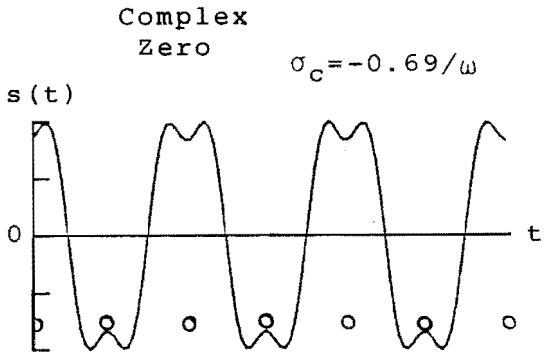


Fig. 2.15.1

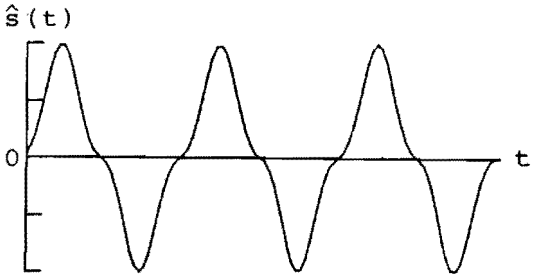


Fig. 2.15.2

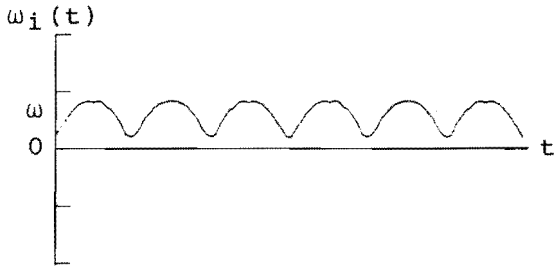


Fig. 2.15.3

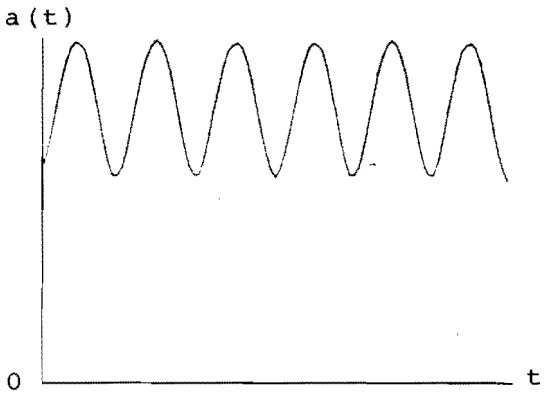


Fig. 2.15.4

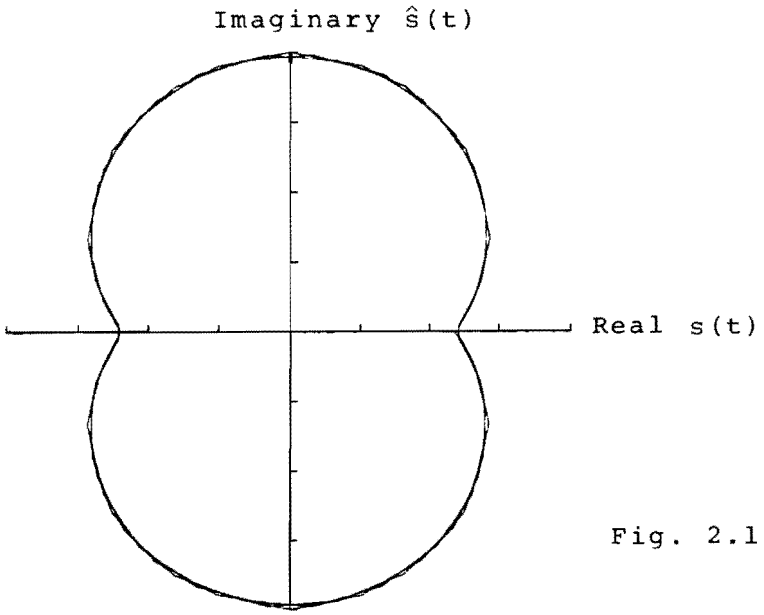


Fig. 2.15.5

Fig. 2.15 Curves for B=0.25

instantaneous waveform fluctuations. The real waveform exhibits complex conjugate pairs of complex zeros ($\sigma_r=\pm 0.78/\omega$) co-located with the non-removed analytic signal zeros in real time.

As predicted, addition of a small third harmonic component has not altered the zero crossing rate or zero crossing positions of s(t). Average instantaneous frequency still corresponds to this zero crossing rate.

Increasing the value of B to 0.5 increases the magnitude of real waveform fluctuations and the size of corresponding instantaneous waveform dips, figure 2.16. As expected, the associated analytic and real signal complex zeros have been moved closer to the real time axis ($\sigma_c=-0.35/\omega, \sigma_r=\pm 0.48/\omega$).

Although s(t) still displays two zero crossings per cycle, the inner loops of vector plot figure 2.16.5 cause $\hat{s}(t)$ to cross zero six times per cycle. As the average instantaneous frequency is still ω radians per second, an estimate of average instantaneous frequency based on the zero crossing rate of $\hat{s}(t)$ would be erroneous.

Further increasing B to 0.9 results in the expected increased magnitudes of instantaneous parameter dips, real waveform fluctuations and vector inner loops, figure 2.17. The associated complex zeros also more closer to the real time axis ($\sigma_c=-0.05/\omega, \sigma_r=\pm 0.17/\omega$).

Once again, s(t) exhibits two zero crossings per cycle while $\hat{s}(t)$ displays six. This situation can be changed, however, by altering the phase relationship between the two components of s(t).

$$s(t)=\cos\omega t-0.9\cos(3\omega t-{}^{2\pi}/_{3}) \qquad \ldots \quad (2.72)$$

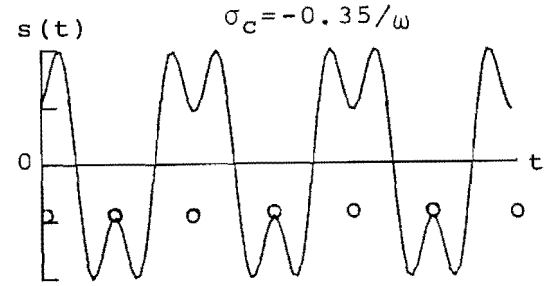


Fig. 2.16.1

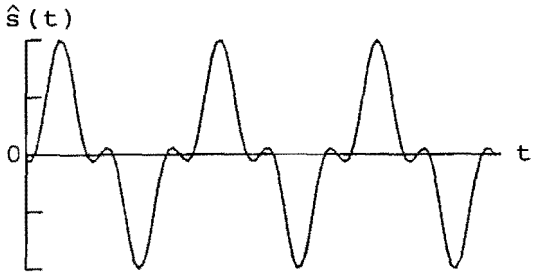


Fig. 2.16.2

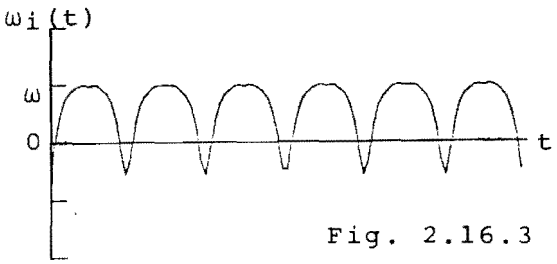


Fig. 2.16.3

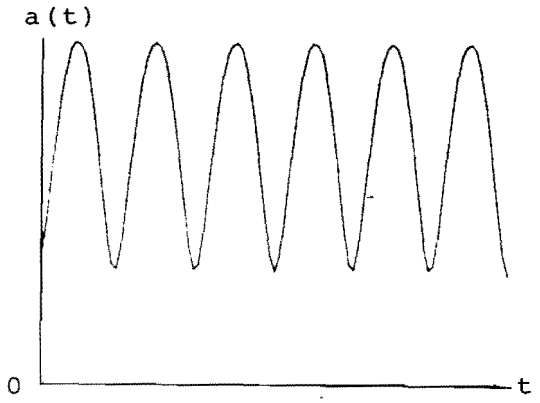


Fig. 2.16.4

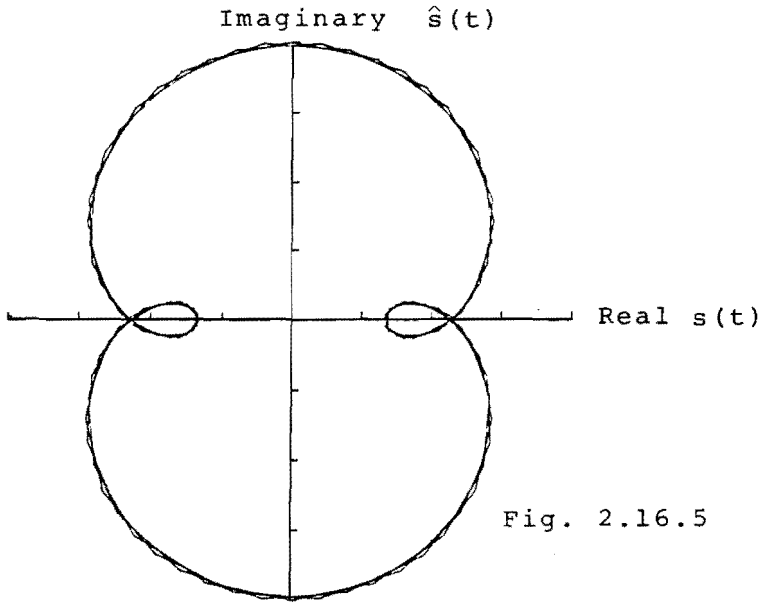


Fig. 2.16.5

Fig. 2.16 Curves for B=0.5

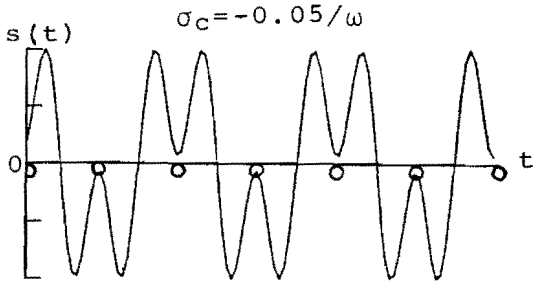


Fig. 2.17.1

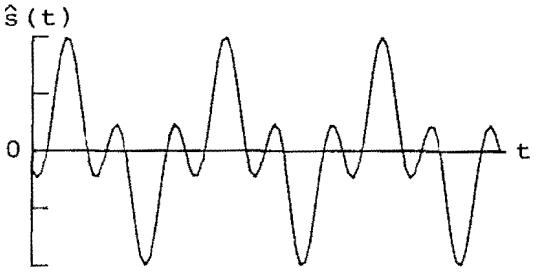


Fig. 2.17.2

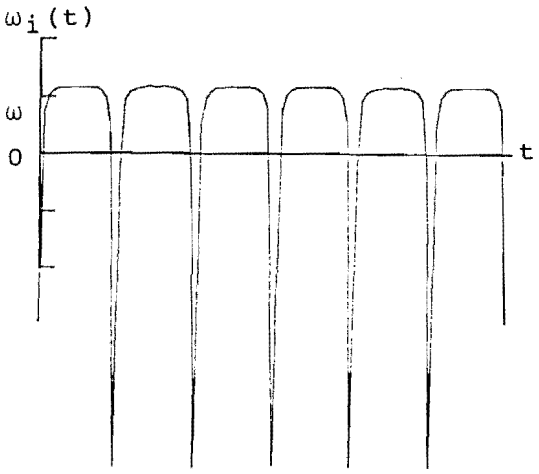


Fig. 2.17.3

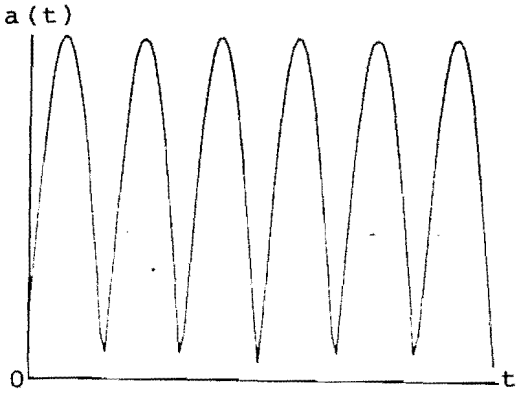


Fig. 2.17.4

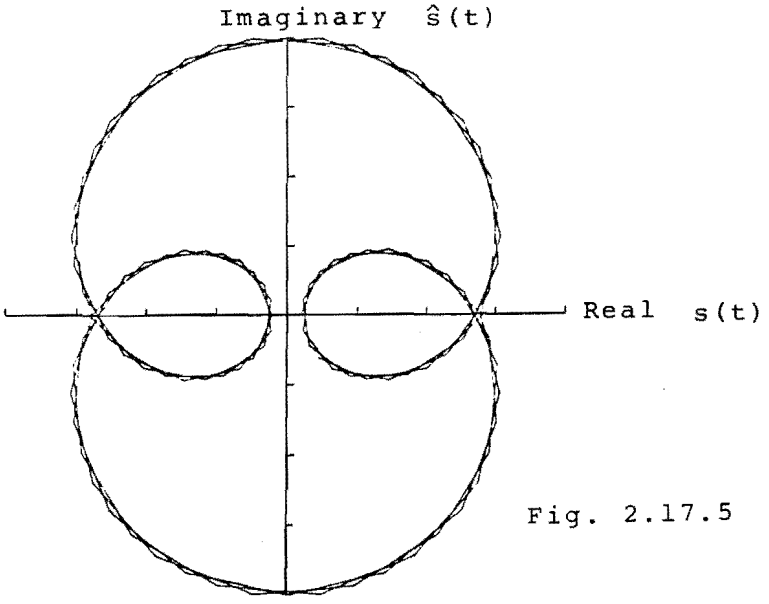


Fig. 2.17.5

Fig. 2.17 Curves for B=0.9

Figure 2.18 shows that this modification does not alter the instantaneous parameter waveforms, except by a time shift, and that the vector plot has been rotated. The real signal s(t) now exhibits six zero crossings per cycle, but average instantaneous frequency is still ω radians per second.

The roots of the analytic signal corresponding to equation (2.72) may be found by factorising

$$e^{j\omega z}\{1-0.9e^{-2j\pi/3}.e^{2j\omega z}\}=0 \qquad \ldots (2.73)$$

It can be seen that σ_c is unaltered by this phase shift.

The roots of the real signal s(z) corresponding to equation (2.72) are defined by

$$-0.9e^{-2j\pi/3}e^{6j\omega z}+e^{4j\omega z}+e^{2j\omega z}-0.9e^{-2j\omega/3}=0 \quad \ldots (2.74)$$

Although equation (2.74) is difficult to solve, the waveforms and vector plot of figure 2.18 indicate that all of the zeros of s(z) are real.

The above exercise shows that LHP analytic signal zeros may translate into real signal fluctuations (complex conjugate pairs of zeros) or pairs of zero crossings. Zero crossings can only be generated if the magnitude of σ_c allows vector inner loops <u>and</u> the correct phase relationship exists between signal components (causing the vector inner loop to intersect the line s(t)=0).

If the two components of equation (2.72) were not phase locked, the real signal zero crossing count would be expected to change continuously between two and six per cycle.

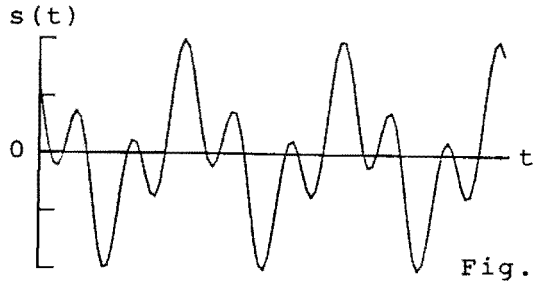


Fig. 2.18.1

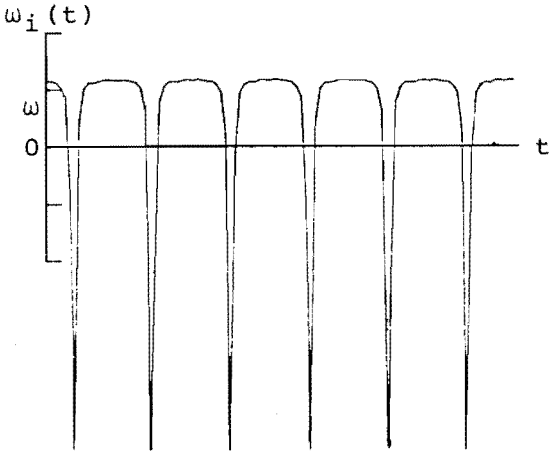


Fig. 2.18.2

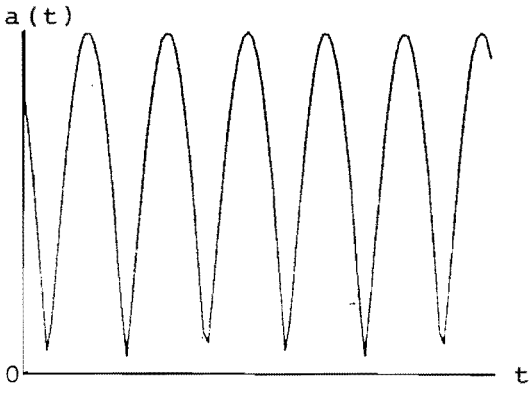


Fig. 2.18.3

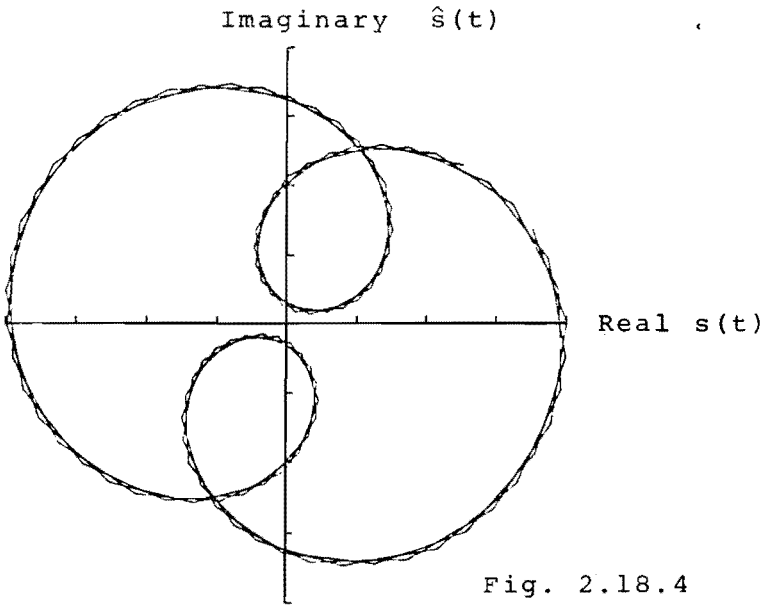


Fig. 2.18.4

Fig. 2.18 Curves for Phase Shifted Third Harmonic

When the third harmonic has greater magnitude than the fundamental, B=2, the real waveform fluctuations become large enough to intersect the time axis, figure 2.19 . Although instantaneous amplitude still displays two dips per cycle, the instantaneous frequency disturbances are now spikes, indicating that the corresponding analytic signal zeros are now UHP (σ_c=+0.35/ω).

Both s(t) and $\hat{s}$(t) exhibit six zero crossings per cycle and examination of the vector plot, figure 2.19.5, reveals that the zero crossing count cannot be altered by rotation. The average instantaneous frequency has now stepped to 3ω and is consistent with the zero crossing rate.

While the lower Fourier component (ω) was dominant, the non-removed analytic signal zeros were LHP and, over one cycle of s(t), these contributed nothing to the average rate of phase change. When the higher component (3ω) became dominant, average instantaneous frequency stepped to 3ω, indicating that each UHP analytic signal zero contributes a phase change of 2π radians per cycle.

The UHP analytic signal zeros of the example always give rise to a pair of real signal zero crossings.

(2.6) INSTANTANEOUS AMPLITUDE - FREQUENCY RELATIONSHIPS

Equations (2.61), (2.62) and the previous example suggest that, under some conditions, ω_i(t) may be recoverable from a(t) and vice-versa. Appropriate conditions can be derived from the equation of an analytic signal

$$\Psi(t)=a(t)e^{j\phi(t)} \qquad \ldots (2.75)$$

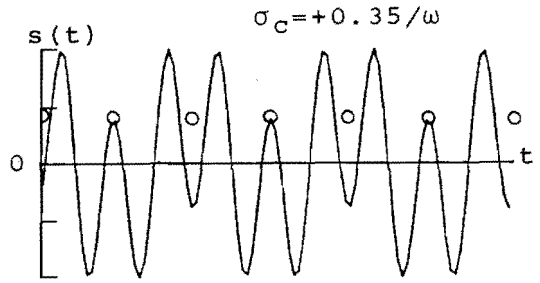


Fig. 2.19.1

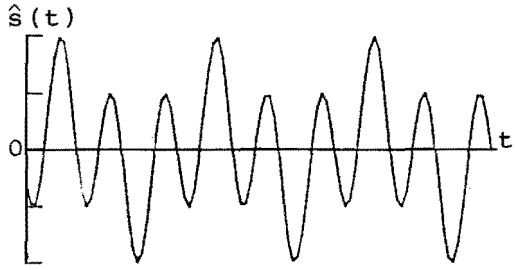


Fig. 2.19.2

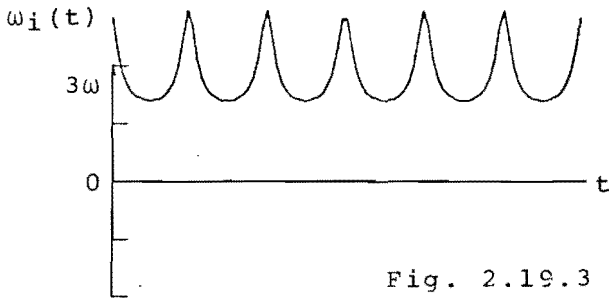


Fig. 2.19.3

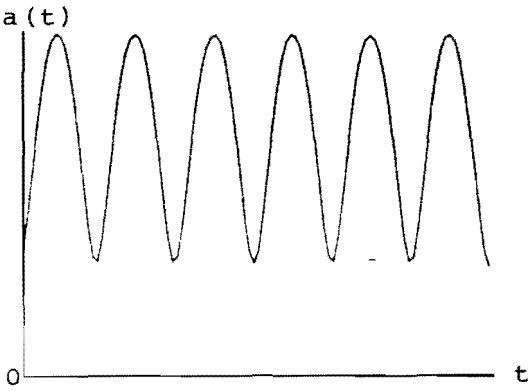


Fig. 2.19.4

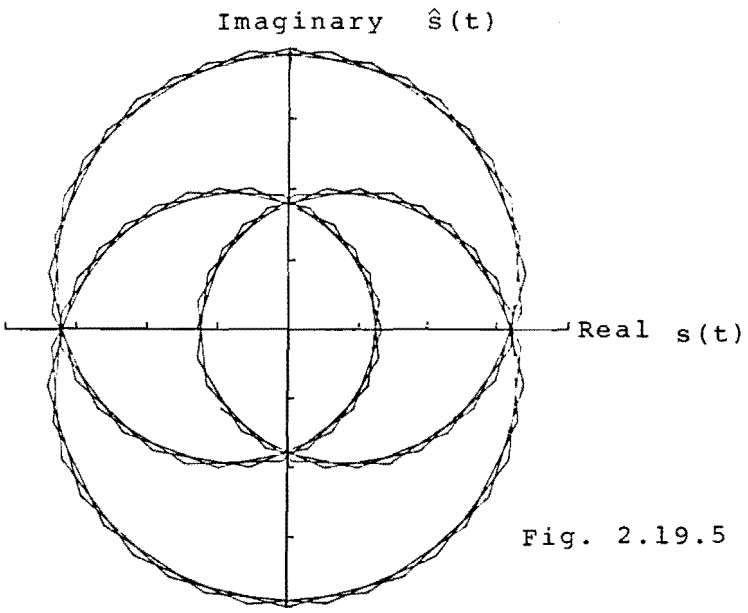


Fig. 2.19.5

Fig. 2.19 Curves for B=2

Taking the natural logarithm of both sides of equation (2.75) gives

$$\ln\{\Psi(t)\}=\ln\{a(t)\}+j\phi(t) \qquad \ldots (2.76)$$

If $\ln\{\Psi(t)\}$ is analytic, then there is a Hilbert transform relationship between the real and imaginary components $\ln\{a(t)\}$ and $\phi(t)$. However, $\ln\{\Psi(t)\}$ can only be analytic (except at infinity) if $\Psi(t)$ has no non-removed UHP zeros. This is known as the Minimum Phase (MP) property (Ref. 86). If $\Psi(t)$ is MP, then

$$\phi(t)=HT\{\ln(a(t))\}$$

or

$$\omega_i(t)=d/dt\{HT\{\ln(a(t))\}\} \qquad \ldots (2.77)$$

The two component signal of the previous example exhibited the MP property when the lower frequency component was dominant. A general bandpass signal, however, will not possess the required analytic zero distribution and must be altered for the relation, equation (2.77), to apply.

When translated to its real projection, an UHP analytic zero always generates a pair of zero crossings. For this reason the MP condition can be generated by precluding zero crossings from the real waveform. Zero crossings can be removed by adding a DC constant k so that

$$g(t)=k+s(t), g(t)>0 \qquad \ldots (2.78)$$

or by exponentiation

$$g(t)=\ell.e^{s(t)} \qquad \ldots (2.79)$$

where ℓ is a positive constant. The analytic function

$$T(t)=g(t)+j\hat{g}(t) \qquad \ldots (2.80)$$

is therefore minimum phase and if $T(t)$ has magnitude $|T(t)|$ and phase $\theta(t)$, we can write

$$g(t)=|T(t)|\cos\{HT\{\ln|T(t)|\}\} \qquad \ldots (2.81)$$

or

$$g(t)=e^{-j\hat{\theta}(t)}\cos\{\theta(t)\} \qquad \ldots (2.82)$$

The equations (2.81) and (2.82) form the basis of compatible single sideband (CSSB) signals which, depending on modulation method, may be demodulated using conventional envelope detectors or phase discriminators. (Ref. 87 - 93)

CHAPTER 3

(3.1) HARDWARE ANALYTIC DECODER FOR SPEECH

The design of equipment to perform simultaneous amplitude and frequency demodulation (analytic decoding) based on equations (3.1) and (3.2)

$$a(t) = (s^2(t)+\hat{s}^2(t))^{\frac{1}{2}} \quad \ldots (3.1)$$

$$\omega_i(t) = {}^{d}/_{dt}\{\tan^{-1}(\hat{s}(t)/(s(t)))\} \quad \ldots (3.2)$$

is greatly influenced by the bandwidth of s(t), the signal to be analysed. For example, the speed at which a digital implementation of equations (3.1) and (3.2) must operate can be minimised by restricting s(t) to the smallest frequency range that will pass an acceptable version of the signal. In the case of speech, this is the telephone bandwidth and any analytic decoder for speech must work over the frequency range 300 Hz $\leq$ f $\leq$ 3400 Hz, or better.

Past research into the parameters of analytic signals and frequency demodulation techniques has spawned hardware which performs a type of analytic decoding (Ref. 94 - 100). Decoders capable of operating in real time usually featured a predominance of analogue circuitry and were often only capable of handling bandwidths amounting to a sub-band of baseband speech. These were used in the analysis of some geophysical signals and animal calls. Computer simulations of such systems make use of Fourier and numerical techniques, and are generally not capable of real time operation.

Basing system design around available technology, the author envisaged a real time decoder which uses analogue techniques in the generation of quadrature signals s(t) and $\hat{s}(t)$, but digital techniques to perform calculations (3.1) and (3.2). Because of time limitations (125 µs between the samples of telephone quality speech sampled

8000 times per second), the squaring operations of equation (3.1) must be performed by memory look-up. Similarly equation (3.2) can be re-written

$$\omega_i(t) = d/dt\{\tan^{-1}(\log^{-1}(\log(\hat{s}(t))-\log(s(t))))\}. \quad . \quad . \quad (3.3)$$

and the division performed by logarithmic look-up and subtraction. The arctan operation may also be performed by look-up.

The resulting analytic decoder design consists of an analogue signal conditioner followed by a digital "pipeline" computer. This is illustrated in figure 3.1.

As indicated in Chapter 2, $\omega_i(t)$ is likely to be unbandlimited and of infinite dynamic range. Naturally, this decoder can only produce an approximation to the instantaneous frequency of speech as the output is bandlimited to 4000 Hz and dynamic range limited to the decoders digital word length.

The design of each section of the analytic speech decoder will be treated separately, and circuit details are presented in Appendix (D).

(3.2) SPEECH CHANNEL PRE-FILTER

Before generation of the orthogonal signal, speech is restricted to the "telephone bandwidth". This is achieved by means of a standard filter design used extensively in our speech research at the University of Canterbury (Appendix (D.1)). The circuit is essentially a six pole low pass filter constructed from three cascaded unity gain Sallen-Key resonators, followed by a four pole high pass section. Figure 3.2 is the resulting frequency response.

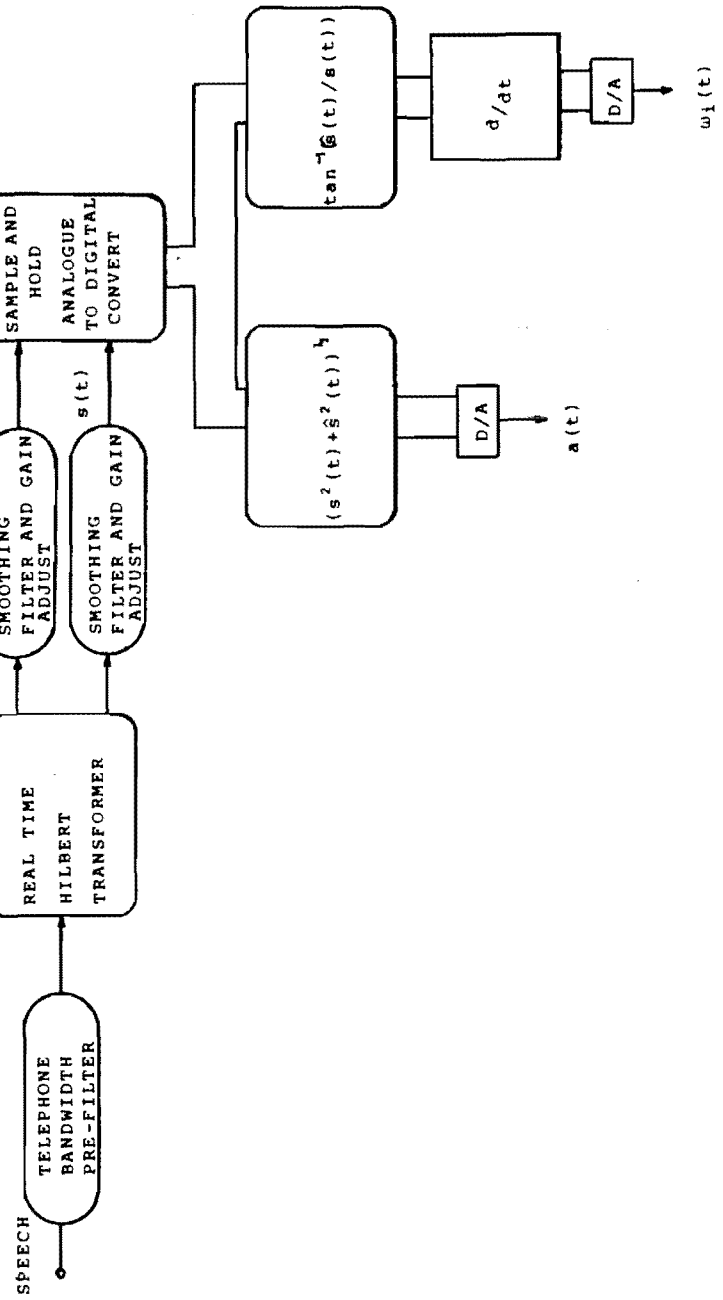


Fig. 3.1 Analytic Decoder

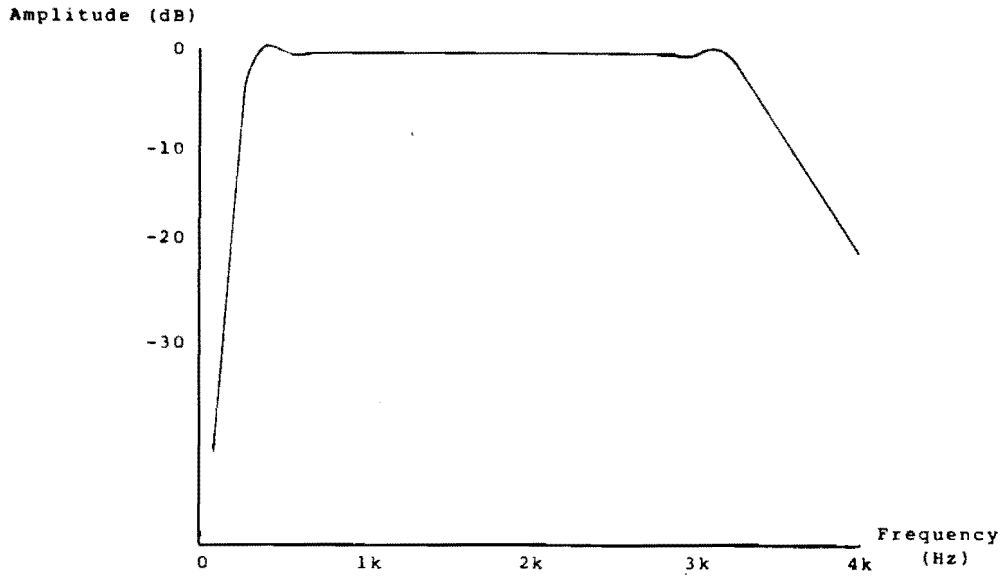


Fig. 3.2 Pre-Filter Frequency Response

(3.3) REAL TIME HILBERT TRANSFORMER

A physical Hilbert transformer can be visualised as a filter with the frequency response

$$H(\omega) = -j.\mathrm{sgn}(\omega) \qquad \ldots (3.4)$$

and corresponding impulse response

$$h(t) = {}^{1}/_{\pi t} \qquad \ldots (3.5)$$

These responses are illustrated in figure 3.3.

Designing a filter to realise the impulse response is difficult as h(t) extends into both positive and negative time. For this reason, a tapped analogue delay line - finite impulse response (FIR) implementation

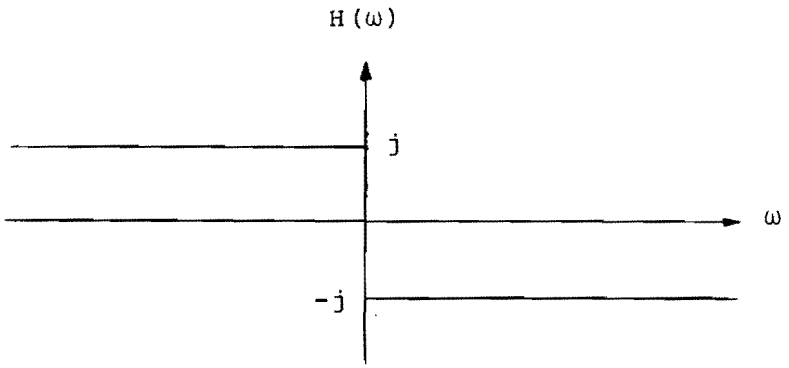


Fig. 3.3.1

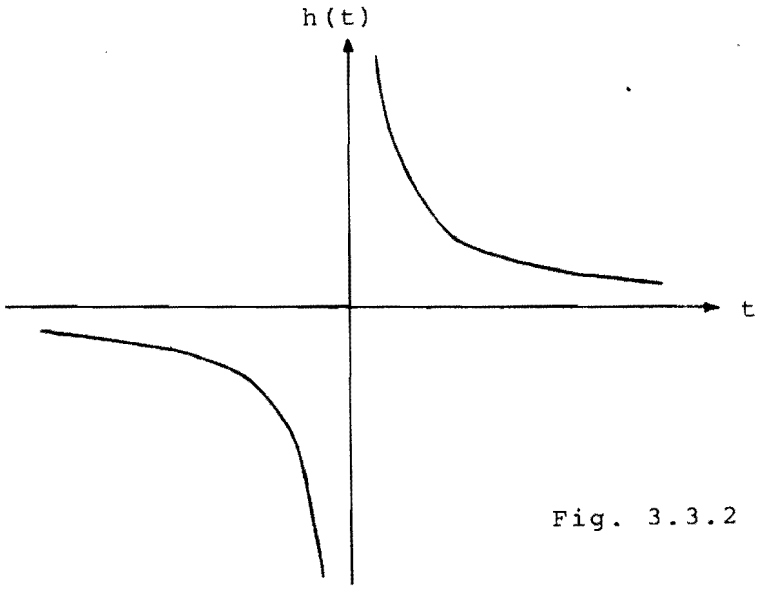


Fig. 3.3.2

Fig. 3.3 Hilbert Transform Frequency and Impulse Responses

presents itself as an attractive possibility.

A tapped analogue delay line (TAD) operates by sampling an input waveform and clocking the analogue samples along the line at the sampling rate. The total delay over an n tap delay line is therefore $n.\tau_s$ where τ_s is the time between samples. If ω_s is the sampling frequency,

$$\tau_s = {}^{2\pi}/_{\omega_s} \qquad \ldots \quad (3.6)$$

A non-causal impulse response may be realised on a TAD by treating the centremost tap as holding the "present" sample, the taps before it as being in negative time and the taps after it as being in positive time. A TAD set up in this type of filter arrangement is illustrated in figure 3.4.

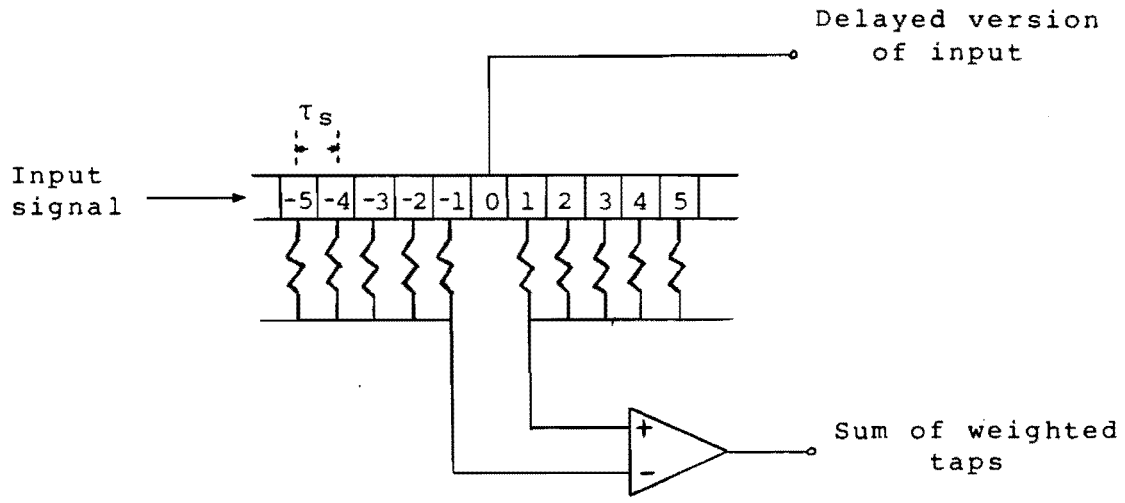


Fig. 3.4 Non-Causal Filter

If the summed tap resistors are weighted by the correct impulse response then the delayed version of the input waveform, taken from tap "0", and the summed output can display a wideband 90° phase difference.

As TAD operation requires that the input be sampled, the correct impulse response will be that of a discrete Hilbert transformer.

The discrete impulse response $h_D(t)$ is found by taking the inverse Fourier transform of the Hilbert transform frequency response restricted by a sampling frequency.

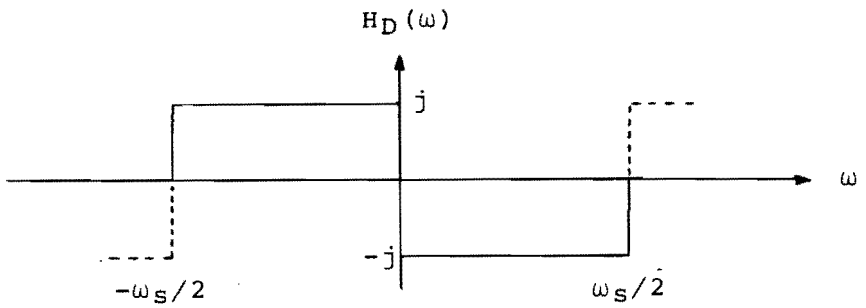


Fig. 3.5 Discrete Hilbert Transform Frequency Response

$$h_D(t) = \int_{-\omega_s/2}^{\omega_s/2} -j.\mathrm{sgn}(\omega).e^{j\omega t}.d\omega \qquad \ldots (3.7)$$

$$= {}^{1}/_{\pi t}(1-\cos({}^{\omega s}/_{2})t) \qquad \ldots (3.8)$$

Rewriting this in terms of τ_s gives

$$h_D(t) = {}^{1}/_{\pi t}(1-\cos({}^{\pi t}/_{\tau_s})) \qquad \ldots (3.9)$$

which is illustrated in figure 3.6. (Ref. 101 - 104).

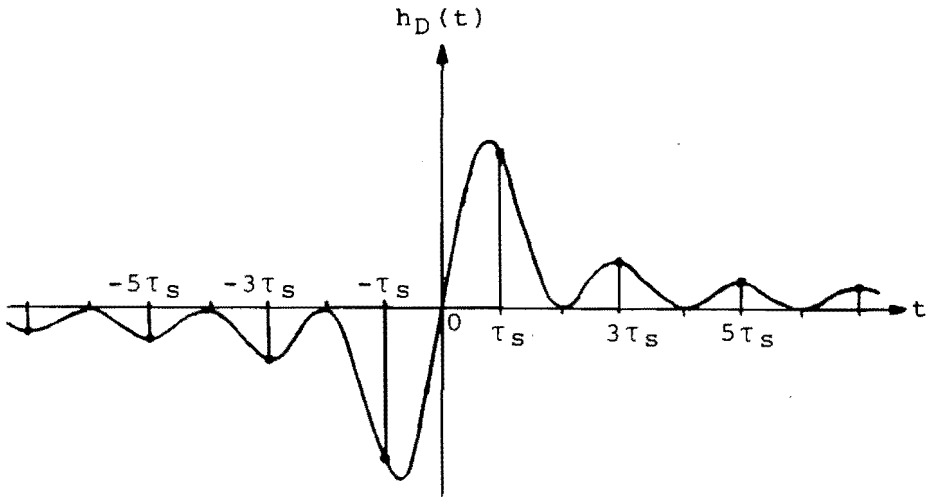


Fig. 3.6 Discrete Hilbert Transform Impulse Response

The tap weightings for a TAD implementation are found by sampling $h_D(t)$ at times corresponding to multiples of τ_s. The weighting values are therefore

$$\ldots h_D(-3\tau_s), h_D(-2\tau_s), h_D(-\tau_s), h_D(\tau_s), h_D(2\tau_s), h_D(3\tau s) \ldots$$

It will be seen from figure 3.6 that every second tap weighting is zero ($h_D(n\tau_s)=0$ when n is an even integer) and the non zero weightings are proportional to ${}^{1}/_{(\pi m\tau_s)}$ where m is an odd integer. Figure 3.7 illustrates this impulse response implemented on an infinite length TAD.

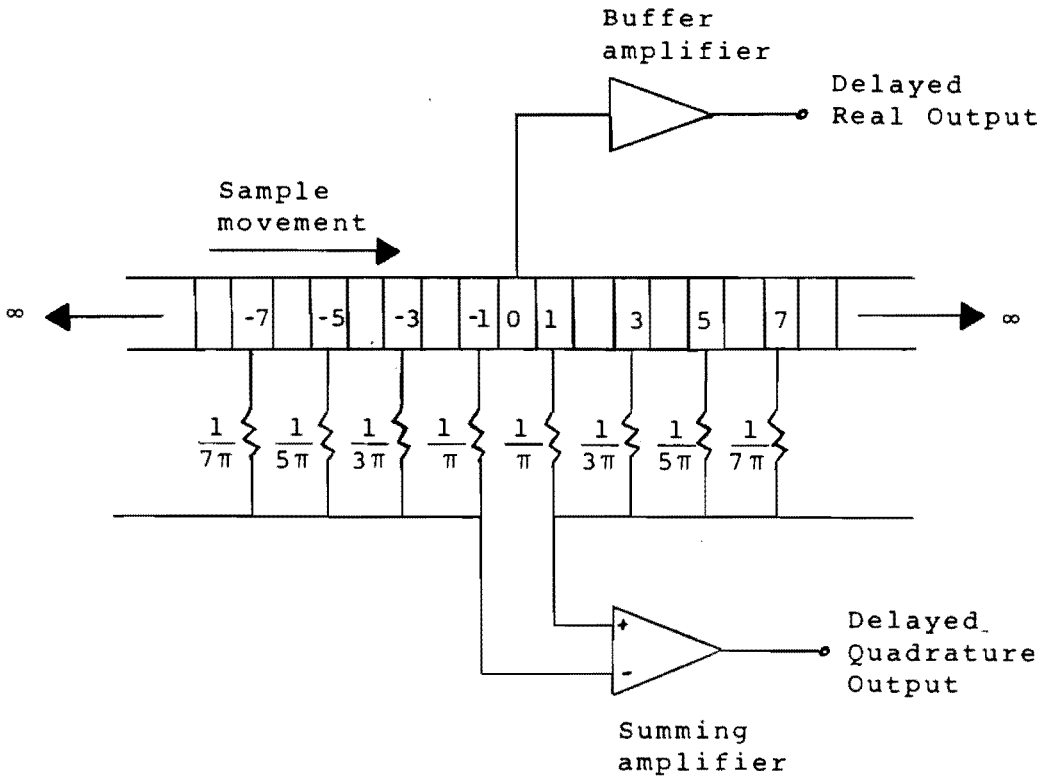


Fig. 3.7 Discrete Hilbert Transformer

Unfortunately, any physical TAD has a finite number of taps and as the author had access to a 32 tap version, it was necessary to limit $h_D(t)$ at $t=\pm 15\tau_s$. Simple truncation of the impulse response at this point would cause unacceptable amplitude ripple on the quadrature frequency response of the discrete Hilbert transformer (Gibbs Phenomenon), and for this reason, $h_D(t)$ was further modified by applying the Gaussian Window of equation (3.10). (Ref. 105)

$$W(t) = \exp(-0.5(5t/30\tau_s)^2) \qquad \ldots (3.10)$$

This windowing function is illustrated in figure 3.8.

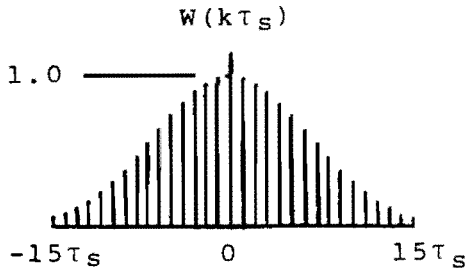


Fig. 3.8 Discrete Gaussian Window

The final practical impulse response was $h_{DW}(t)$, where

$$h_{DW}(t) = {}^{1}/_{\pi t}(1-\cos({}^{\pi t}/_{\tau s})).\exp(-0.5(5t/30\tau s)^2), \tau s = 125\mu s \quad (3.11)$$

and this was tested by computer simulation. The tests yielded the frequency response of figure 3.9 and as this appeared suitable for use with telephone quality speech, a hardware version was built and tested, Appendix (D.2).

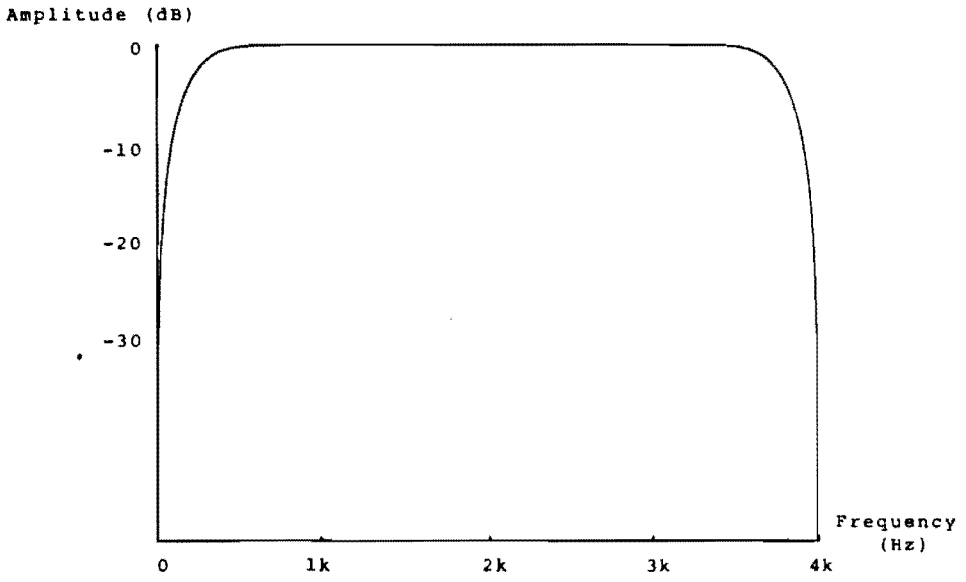


Fig. 3.9 Simulation Frequency Response

Some problems were encountered in hardware when setting small tap weightings with "ten-turn trimpots", and it was discovered that only the first five non zero weightings on either side of t=0 were significantly greater than zero. This is illustrated in figure 3.10 which is an oscilloscope photograph of $h_{DW}(k.\tau_s)$.

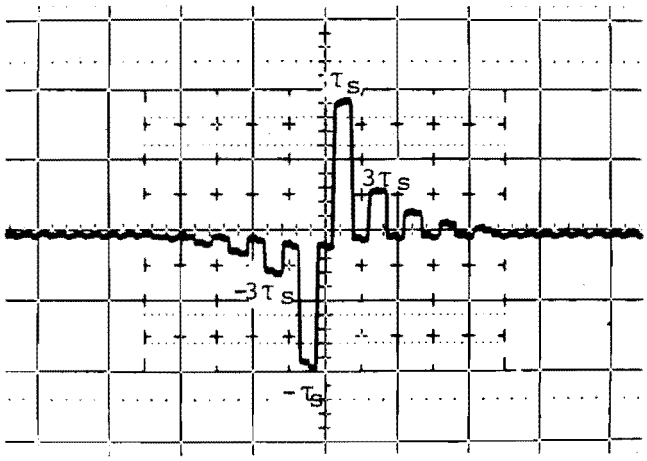


Fig. 3.10 Practical Impulse Response

Being unable to set the remaining taps made little difference to the final frequency response which is illustrated in figure 3.11. It can be seen that this agrees well with the simulation.

Proper operation of the Hilbert transformer was confirmed by generating simple vector patterns through real and quadrature modulation of the "x" and "y" axes of an oscilloscope display. These conformed to the theory of Chapter 2 and are presented later with the corresponding time waveforms.

A second hardware implementation of a real time Hilbert transformer was attempted using an Intel 2920

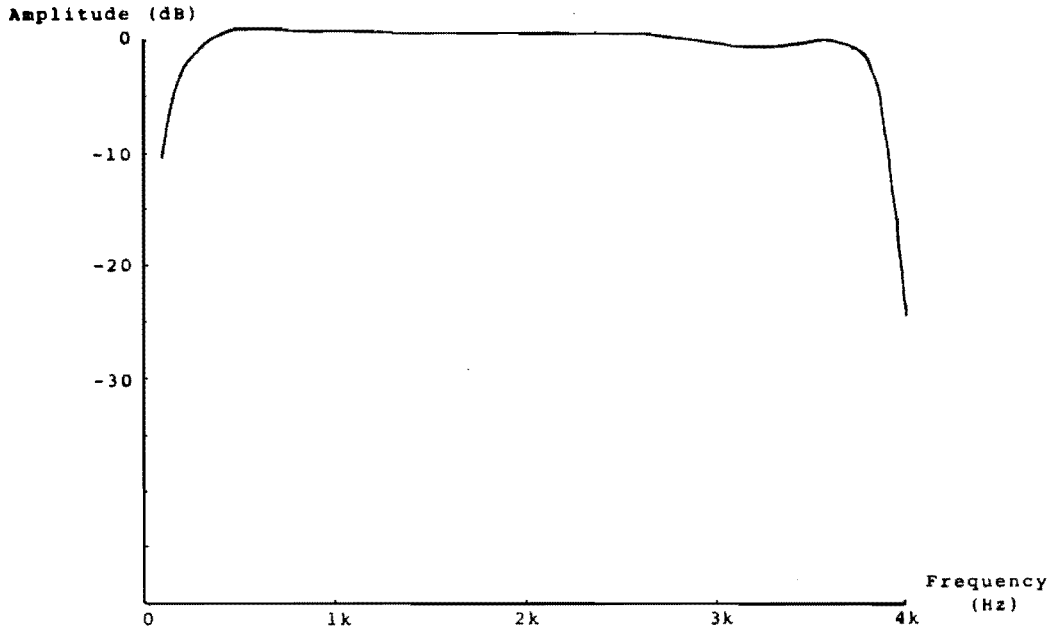


Fig. 3.11 Hilbert Transformer Frequency Response

signal processing microprocessor. Essentially, the 2920 was programmed to emulate a delay line, with tap weighting performed by digital shift and add type multiplication. Although the microprocessor had sufficient computing power, problems with the development kit prevented the successful completion of this version.

(3.4) REAL AND QUADRATURE SMOOTHING FILTERS

Delayed real and quadrature signals generated by the TAD FIR filter are output in the form of sampled and held pulse amplitude modulated waveforms. In order to obtain the smooth analogue waveforms required for display and further processing, it was necessary to re-filter. Although this could be performed using two speech pre-filters (Section 3.2), it must be noted that for quadrature relationships to be maintained, both smoothing filters must

have identical amplitude and phase responses.

Construction of two identical bandpass filters is obviously impossible, but the ideal was approached using hand picked and low tolerance components in the design of Section 3.2. The resulting filters have the frequency response of figure 3.2 and when properly adjusted, track well in phase. The only circuitry change to the design of Section 3.1 is the addition of gain adjustment stages. Appendix (D.3).

The filtered signals, $s(t)$ and $\hat{s}(t)$, are telephone quality voice and quadrature voice suitable for display, sampling and further processing.

(3.5) DIGITAL DATA ACQUISITION

Figure 3.12 is a block diagram of the sampling and analogue to digital conversion circuitry, including a timing diagram of the associated control waveforms.

The quadrature relationship between waveforms is maintained by simultaneous sampling of $s(t)$ and $\hat{s}(t)$. If the sampling occurs at time $t=k\tau_s$ (where k is an integer) then the quadrature samples $s(k\tau_s)$ and $\hat{s}(k\tau_s)$ become available for analogue to digital (A/D) conversion.

Through the multiplex circuitry, $\hat{s}(k\tau_s)$ is selected for the first A/D conversion. The conversion is performed to 8 bits accuracy and a digital representation of $\hat{s}(k\tau_s)$ appears at the convertors output within 4μs. This value is latched further along the digital chain before $s(k\tau_s)$ is selected and converted. The digital representation of $s(k\tau_s)$ is latched at the converter output for the remainder of the clock cycle.

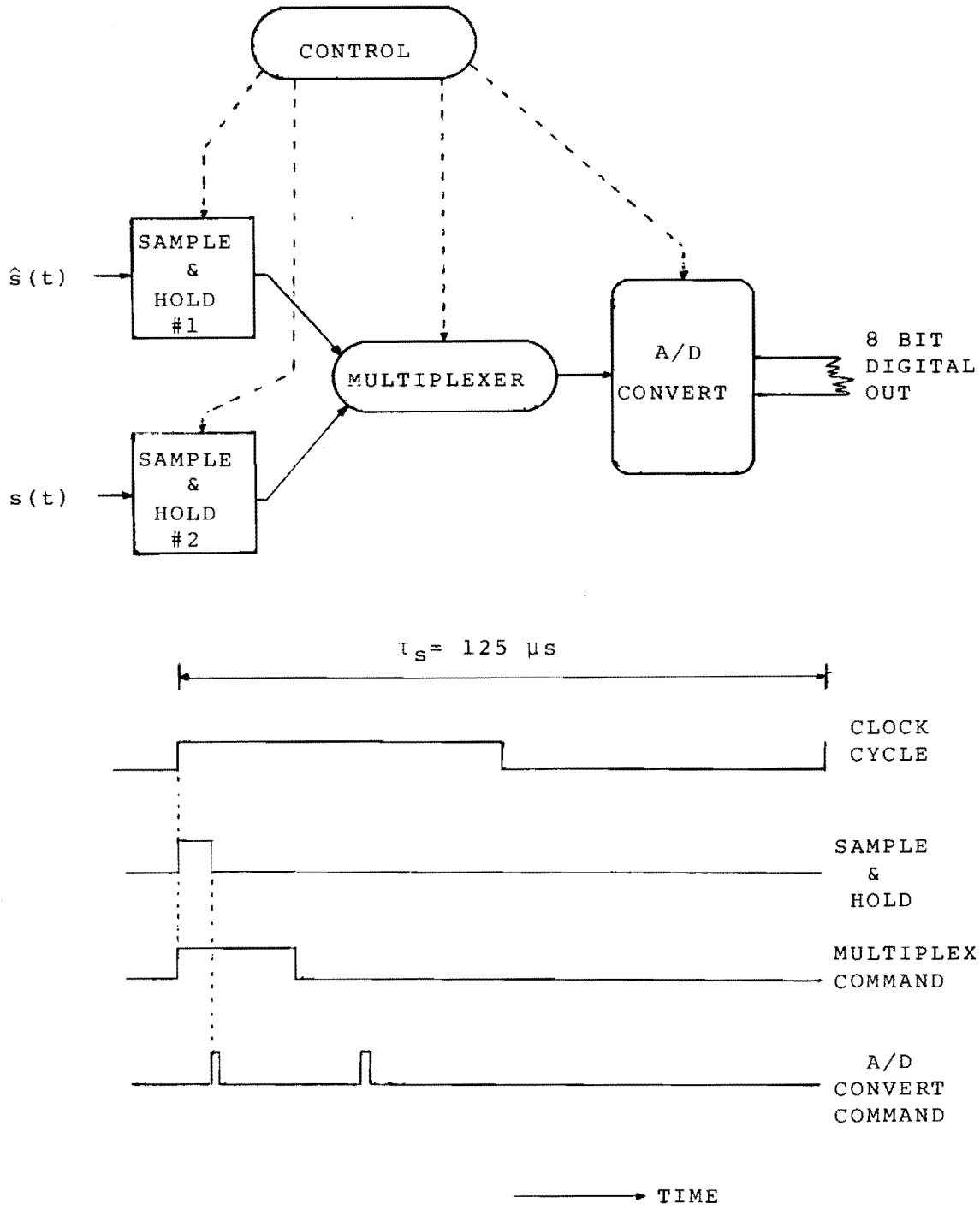


Fig. 3.12 Data Acquisition Circuitry

Although this conversion could have been performed to 12 bits accuracy (over a slightly longer conversion time), 8 bits were deemed sufficient for an approximate representation of $\omega_i(t)$. The use of this shorter digital word also greatly simplified the design of the remaining digital hardware.

As stated previously, the division of equation (3.2) is to be carried out by logarithmic subtraction. However, for successful logarithmic conversion, the variables to be converted must never become negative or zero.

After A/D conversion, both $\hat{s}(k\tau_s)$ and $s(k\tau_s)$ would normally appear as "offset binary" digital words, for which form it is simple to obtain a separate sign - magnitude representation. Removing the sign bit would prevent the variables from becoming negative but either $\hat{s}(k\tau_s)$ or $s(k\tau_s)$ could still be zero. This problem is overcome by changing the A/D output format to "modified offset binary". (Ref. 104).

The format of modified offset binary is illustrated in Table 3.1.

Modified Offset Binary	Equivalent Scaled Value
11111111	+255
11111110	+253
- -	-
10000001	+3
10000000	+1
- - - - - - - - - - - - - - - -	Zero
01111111	-1
01111110	-3
- -	-
00000001	-253
00000000	-255

Table 3.1

The zero level is set to half way between "10000000" and "01111111" by adjusting the A/D input D.C. offset until a grounded input causes the output to flicker between the two digital words. In the unmodified offset case, "10000000" represents zero.

Scaled decimal representations are used for convenience to avoid "10000000" becoming +0.5, "10000001" becoming +1.5 and so on.

Circuitry details of the data acquisition and conversion system are presented in Appendix (D.4).

(3.6) INSTANTANEOUS AMPLITUDE

Figure 3.13 is a block and timing diagram of the digital hardware for calculating instantaneous amplitude. It is a direct implementation of equation (3.1) rewritten

$$a(kT_S) = \left(s^2(kT_S) + \hat{s}^2(kT_S)\right)^{\frac{1}{2}} \quad \ldots \quad (3.12)$$

To ensure sufficient speed of operation, all logic is transistor - transistor logic (TTL) or low power Schottky TTL and the programmable read only memories (PROMS) are 256 x 8 bit fusible link bipolar types. Circuitry details are contained in Appendix (D.5).

Once the first conversion is complete, $\hat{s}(kT_S)$ is latched at the A/D output and the digital word is allowed to ripple through the "modified offset binary to magnitude" conversion logic and the squaring PROM (look up table). This process is reasonably quick, and when sufficient time has elapsed, $\hat{s}^2(kT_S)$ is latched for the remainder of the clock cycle.

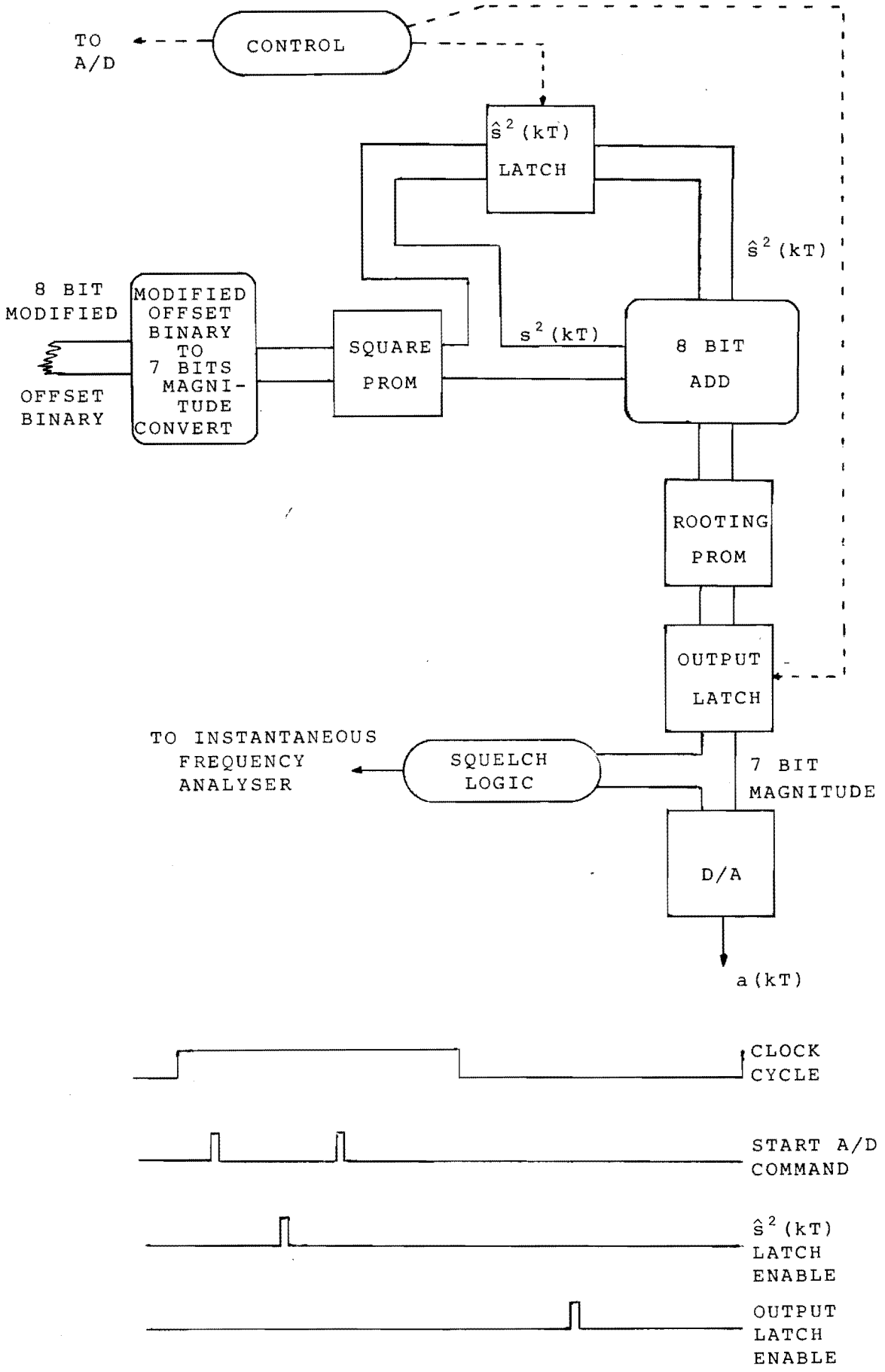


Fig. 3.13 Instantaneous Amplitude Circuitry

When $s(k\tau_s)$ becomes availabe, it too ripples through the conversion logic and squaring PROM. $s^2(k\tau_s)$ is added to the stored $\hat{s}^2(k\tau_s)$ and the rooted sum is eventually latched at the input of a digital to analogue (D/A) converter. Instantaneous amplitude $a(k\tau_s)$ appears as an analogue voltage at the D/A output.

Only the 7 bit sample magnitudes are involved in the squaring operation as the sign bits are of no consequence. The 8 bit output of the squaring look up table represents the most significant 8 bits from the dynamic range of $\hat{s}^2(k\tau_s)$ and $s^2(k\tau_s)$, and this format is taken into account during the square rooting operation. As the square root look up table is essentially the "inverse" of the squaring PROM, $a(k\tau_s)$ appears as a 7 bit magnitude. This is consistent with instantaneous amplitude being a positive function of time.

The squelch logic is based on the least significant bits of $a(k\tau_s)$ and is provided for use when making chart recordings of instantaneous frequency.

(3.7) INSTANTANEOUS PHASE

Figure 3.14 is the block and timing diagram of equipment to perform the calculation

$$\phi_i(k\tau_s) = \tan^{-1}\{\hat{s}(k\tau_s)/s(k\tau_s)\} \quad \ldots (3.13)$$

Once again, all logic is TTL or Schottky TTL and PROMs are bipolar fusible link. Circuitry details are in Appendix (D.6).

The instantaneous phase circuit shares the "modified offset binary to 7 bit magnitude plus sign" conversion logic, except that in this case the sign bit is not discarded. When $\hat{s}(k\tau_s)$ is available, it ripples through the conversion logic and its sign bit is latched for future reference. The magnitude of $\hat{s}(k\tau_s)$ carries on through the

THE LIBRARY
UNIVERSITY OF CANTERBURY
CHRISTCHURCH, N.Z.

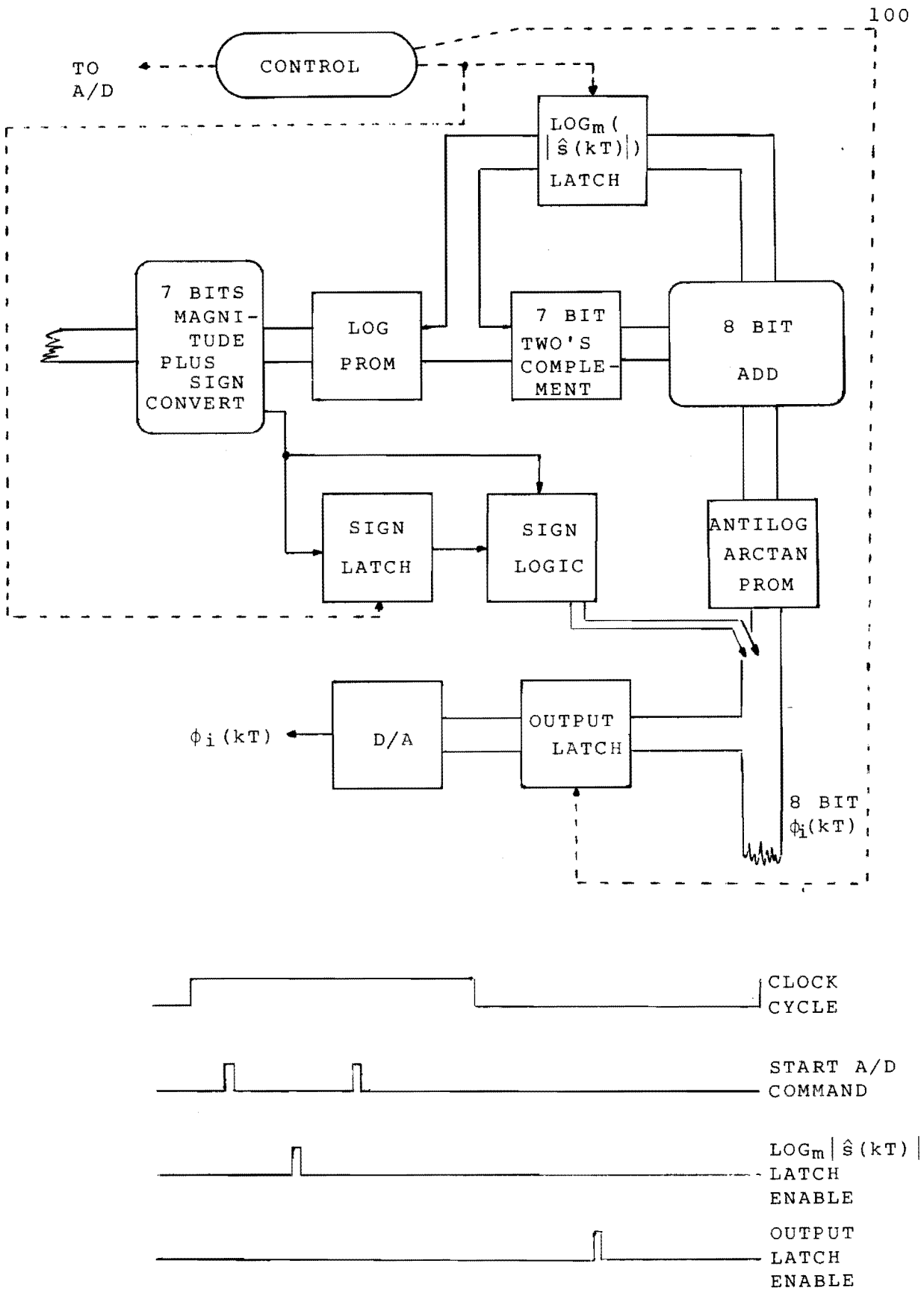


Fig. 3.14 Instantaneous Phase Circuitry

log look up table and log $m|\hat{s}(kT_s)|$ is latched.

The sign bit of $s(kT_s)$ is also removed and log $m|s(kT_s)|$ undergoes a two's complement operation before being added to the latched value. The resulting sum can be written

$$\text{Sum } (kT_s) = \log m|\hat{s}(kT_s)| - \log m|s(kT_s)| \quad . . \quad (3.14)$$

This is passed through a combined antilog and arctan look up table which outputs a 6 bit angle magnitude. The angle is resolved into one of the four possible quadrants by recombining with the processed sign bits of $\hat{s}(kT_s)$ and $s(kT_s)$. The result is an 8 bit representation of $\phi_i(kT_s)$.

To ensure that the entire dynamic range of the log look up table is utilised a logarithmic conversion base m is used. For full utilisation

$$\log m(255) = 127 \quad . . . \quad (3.15)$$

where 255 is the largest scaled sample magnitude and 127 is the largest 7 bit PROM output. Rearranging equation (3.15)

$$m = (255)^{1/127} \quad . . . \quad (3.16)$$

Antilog and arctan processes have been combined in a single look up table with the antilog being the "inverse" of the log process. The angle output, when combined with the two sign bits, can be displayed by means of a latch and D/A converter. It is a bipolar signal in the range $-\pi \leq \phi_i(kT_s) \leq \pi$.

(3.8) INSTANTANEOUS FREQUENCY

Processing instantaneous phase to become instantaneous frequency involves a differentiation with respect to time. When instantaneous phase is represented by a regular sequence of digital words, however, the differentiation becomes a

matter of differencing the present and most recent phase samples according to the equation

$$\omega_i(k\tau_s) = \phi_i(k\tau_s) - \phi_i((k-1)\tau_s) \quad \ldots (3.17)$$

Equipment to perform this differencing is illustrated in figure 3.15 along with the associated timing diagram. Circuit details are in Appendix (D.7).

Latch timing is arranged so that the previous phase value $\phi((k-1)\tau_s)$ is held in the "previous value latch" while the present value $\phi(k\tau_s)$ is permitted to enter the "present value latch". Once again, a subtraction is performed by two's complementing and adding and the sum is fixed by the output latch. Before a new phase value $\phi_i((k+1)\tau_s)$ is allowed to enter the "present value latch", $\phi_i(k\tau_s)$ is transferred to the previous value latch, thus permitting the next differencing operation.

From the analyses of Chapter 2, it is known that instantaneous phase slopes may become negative. As the phase waveform is limited to the range $-\pi \leq \phi_i(k\tau_s) \leq \pi$, a small negative increment of phase could be confused with a large positive increment. This is illustrated in figure 3.16.

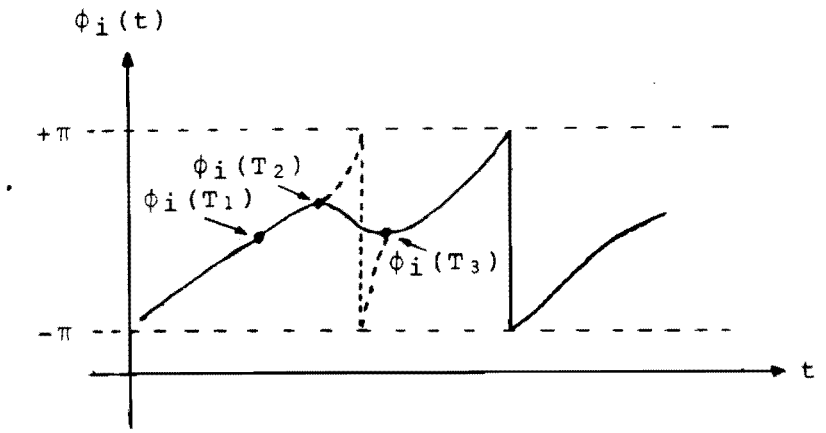


Fig. 3.16 Possible Phase waveform

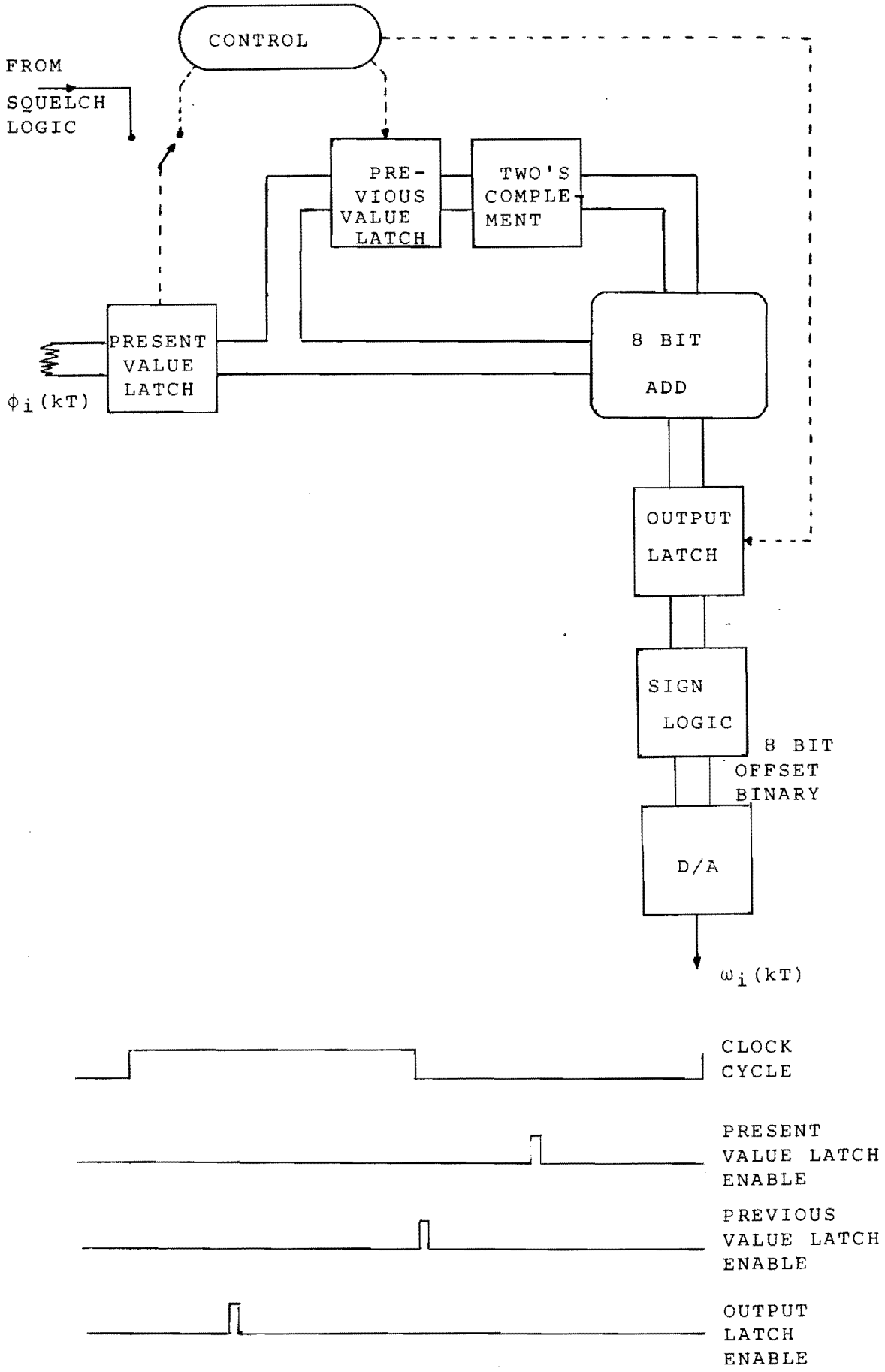


Fig. 3.15 Instantaneous Frequency Circuitry

The instantaneous frequency calculated from phase values at T_1 and T_2 is

$$\omega_i = \phi_i(T_2) - \phi_i(T_1) \quad \ldots (3.18)$$

and in this case ω_i is small and positive. For the next differencing operation,

$$\omega_i = \phi_i(T_3) - \phi_i(T_2) \quad \ldots (3.19)$$

ω_i is ambiguous, resulting from either a small negative phase slope or large (almost 2π) phase increment.

The ambiguity is resolved by the sampling theorem which limits the maximum allowable phase increment between adjacent samples to π radians. A detected increment of greater than π radians therefore indicates that the phase slope is actually negative.

The sign logic following the latch in figure 3.15 converts erroneous large positive phase differences into negative phase differences by subtracting 2π radians from any phase step greater than π.

When viewing the display of $\omega_i(k\tau_s)$ it should be noted that the waveform is delayed by about one clock cycle with respect to $a(k\tau_s)$. This is due to the additional differentiation processing.

When in use, the squelch logic (figure 3.15) disables the "present value latch" when no significant instantaneous amplitude is detected. During the silence between words, this has the effect of reducing $\omega_i(k\tau_s)$ to zero as, after one clock cycle, both the previous and present value latches will hold the same phase value.

(3.9) OPERATION AND DISPLAY

Display of instantaneous waveforms is in the form of cathode ray oscilloscope photographs or high speed chart recordings. Photographs are suitable for displaying fine structure over a few wavelengths while the necessarily low pass nature of chart recordings makes them useful for indicating average values over many wavelengths.

(3.9.1) SYSTEM TESTS

The operation and linearity of the system was verified using a sinusoidal test signal of variable amplitude and frequency.

Figure 3.17.1 is a photograph of s(t) and $\hat{s}(t)$ originating from a 900 Hz test tone. The signals appear to be in perfect quadrature and this is verified by the circular nature of the vector locus figure 3.17.2.

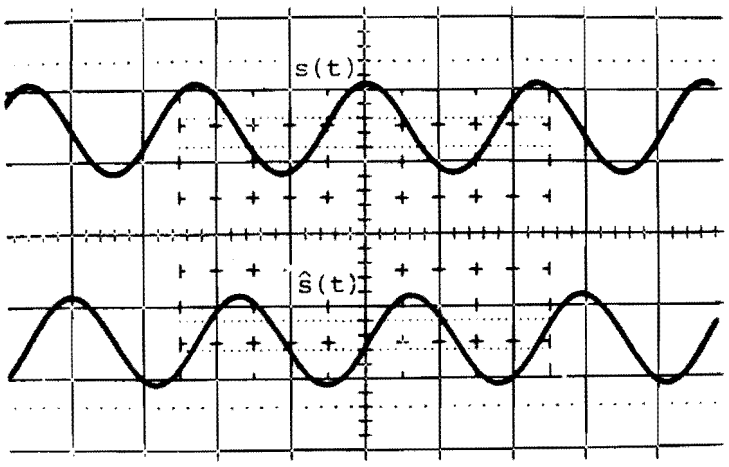


Fig. 3.17.1 Real and Quadrature Waveforms

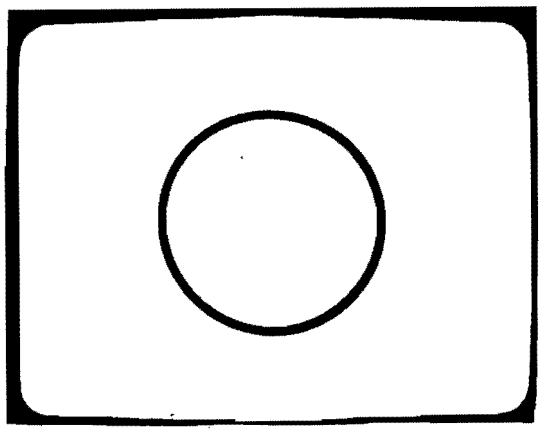


Fig. 3.17.2 Vector Locus

For the same test tone, figure 3.18.1 is the display of instantaneous amplitude and frequency. As expected, these are constant with respect to time, except for granular noise visible on $\omega_i(kT)$.

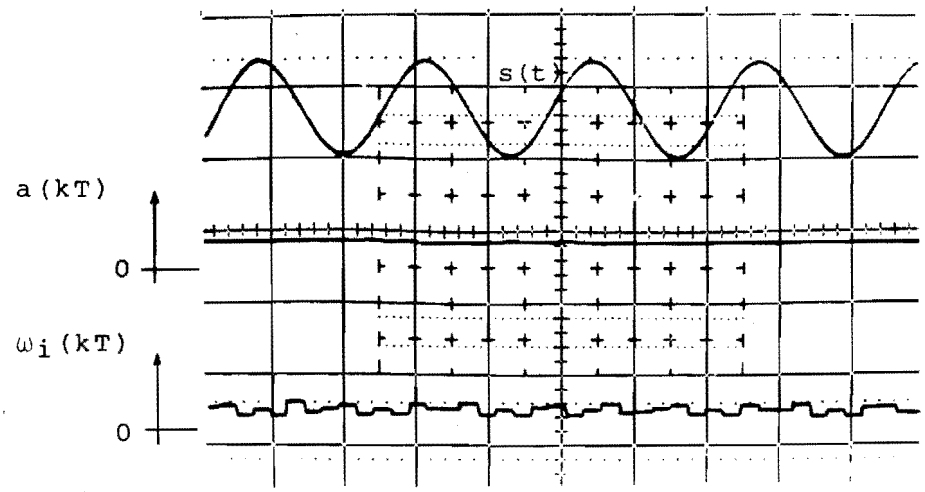


Fig. 3.18.1 Instantaneous Amplitude and Frequency Waveforms

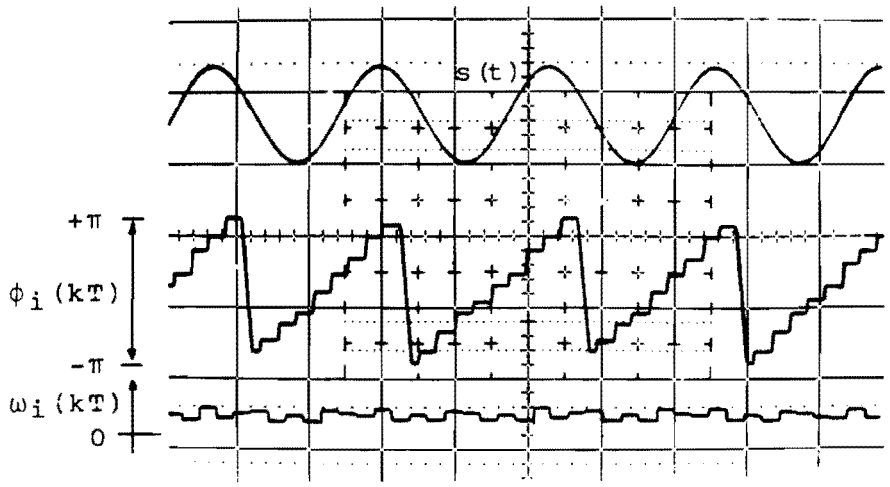


Fig. 3.18.2 Instantaneous Phase & Frequency Waveforms

Figure 3.18.2 illustrates the phase function $\phi_i(kT)$ ramping between $-\pi$ and π. Absolute phase should not be related to the s(t) reference waveform as this display is derived before the digital processing delay. Phase changes, however, can be related to $\omega_i(kT)$.

Figures 3.19.1 and 3.19.2 are plots of a(kT) readings from a constant frequency amplitude sweep and $\omega_i(kT)$ from a constant amplitude frequency sweep respectively. These curves, and observations of vector plot distortion at high frequency, indicate that average $\omega_i(kT)$ tracks a sinusoidal frequency linearly up to 3 kHz. As the bandpass filtering ensures little voice energy above this frequency, the resulting distortion was considered to be of little consequence.

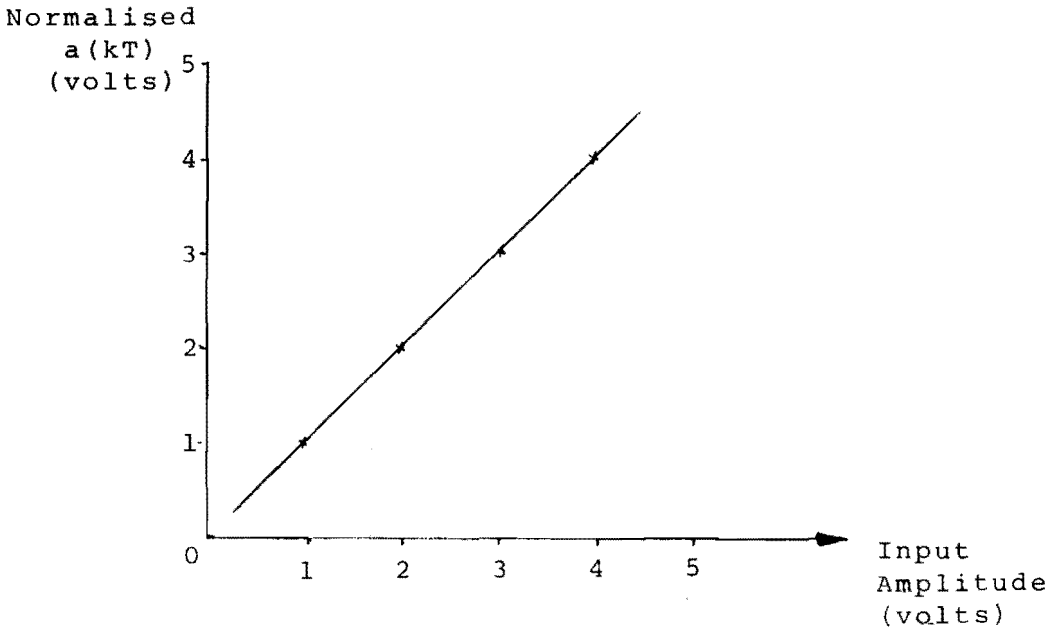


Fig. 3.19.1

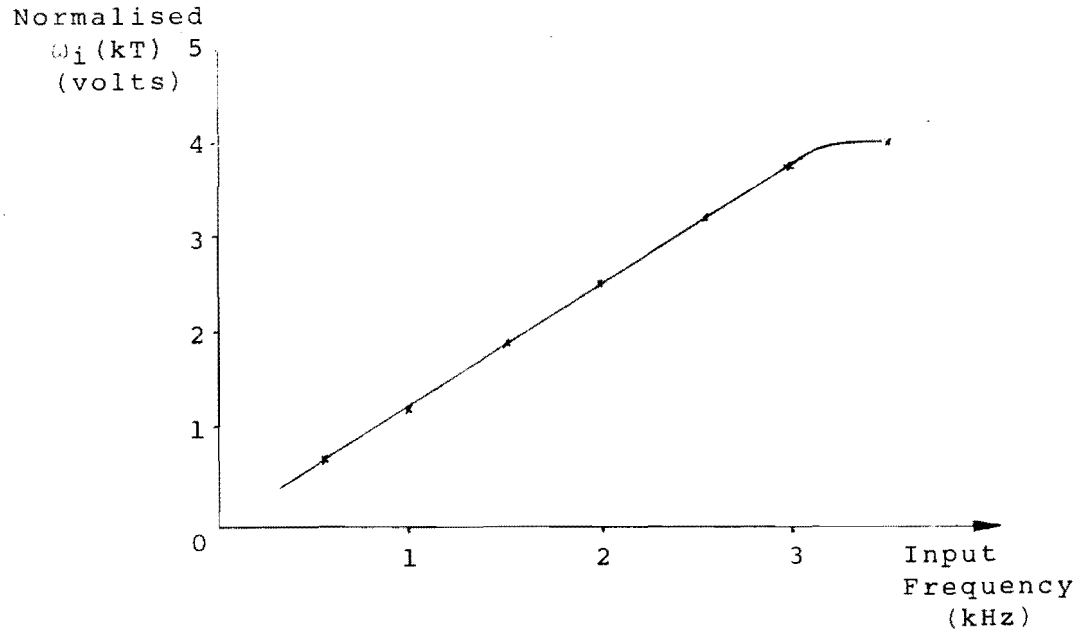


Fig. 3.19.2

Fig. 3.19 Normalised Instantaneous Amplitude and Frequency Responses

To test system operation under all conditions it was necessary to apply a waveform known to generate negative instantaneous frequencies. This test was performed using a signal consisting of a variable amplitude 800 Hz fundamental and second harmonic. The three conditions to be investigated were:

(a) sufficient second harmonic to cause a periodic slowing of the phase ramp with the associated instantaneous amplitude and frequency dips,

(b) larger second harmonic causing negative phase slopes, large instantaneous amplitude dips and negative instantaneous frequencies, and

(c) dominant second harmonic resulting in a doubling of the average rate of phase change, periodic instantaneous amplitude dips and corresponding instantaneous frequency rises.

Figure 3.20.1 is the vector plot of the test signal satisfying condition (a). The expected dips of instantaneous amplitude and frequency can be seen in figure 3.20.2 and figure 3.20.3 shows the correspondence between periodic slowing of the phase ramp and instantaneous frequency dips.

The inner loop of vector plot figure 3.21.1 indicates that the second harmonic is of sufficient amplitude to cause negative instantaneous frequencies and thus satisfy condition (b). Figures 3.21.2 and 3.21.3 show the predicted large instantaneous amplitude dips, negative phase slopes and periodic negative instantaneous frequency excursions.

Making the second harmonic component of larger amplitude than the fundamental causes the inner loop of vector plot figure 3.22.1 to encompass the origin.

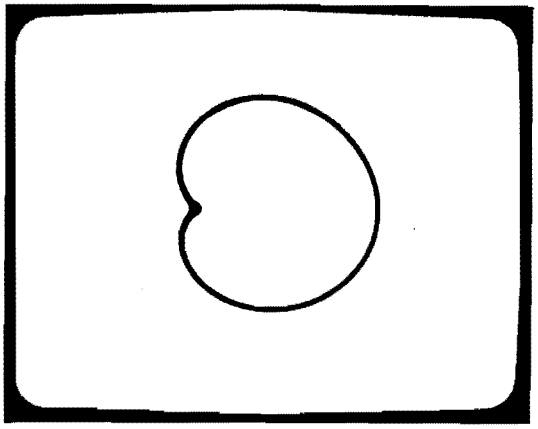


Fig. 3.20.1

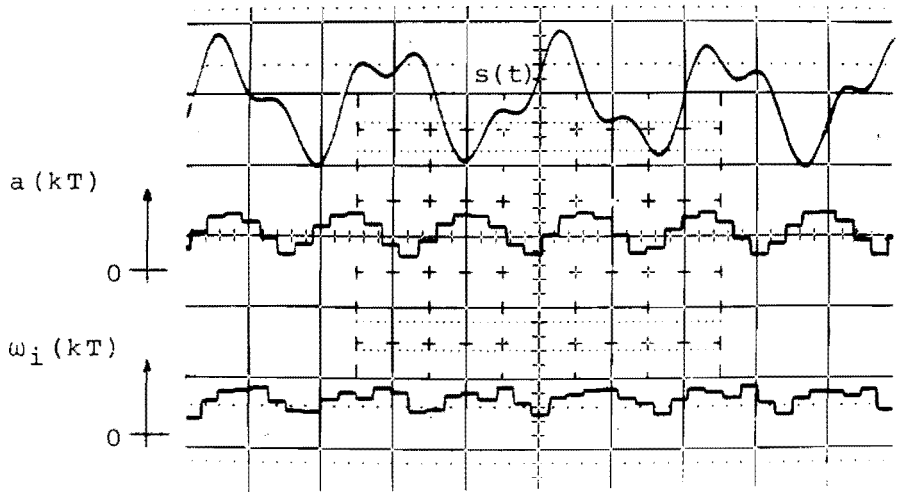


Fig. 3.20.2

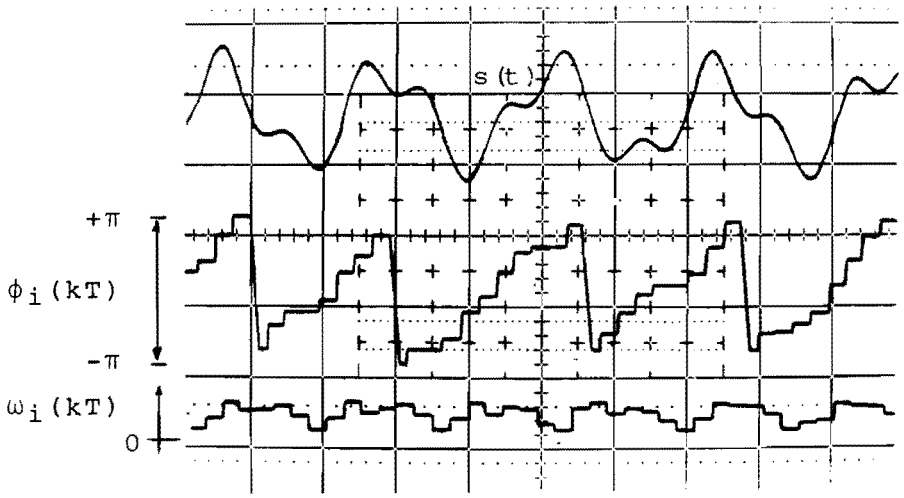


Fig. 3.20.3

Fig. 3.20 Sinusoid Plus Small Second Harmonic

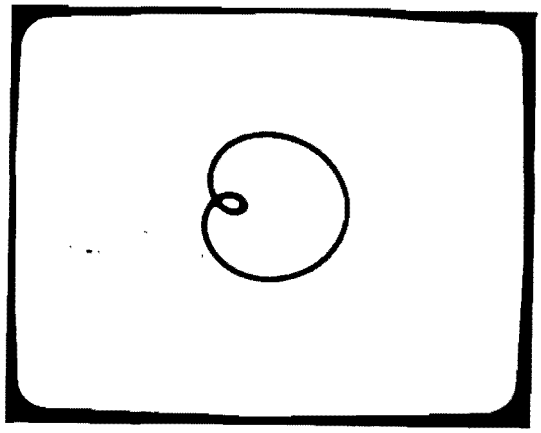


Fig. 3.21.1

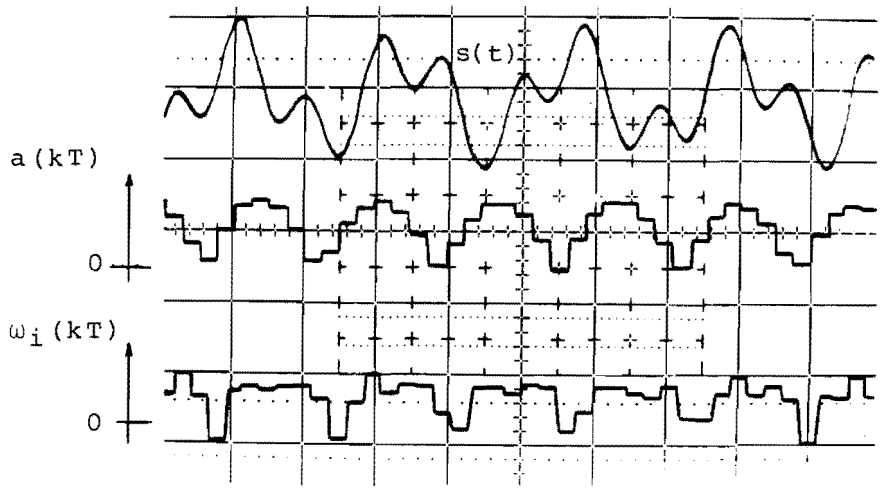


Fig. 3.21.2

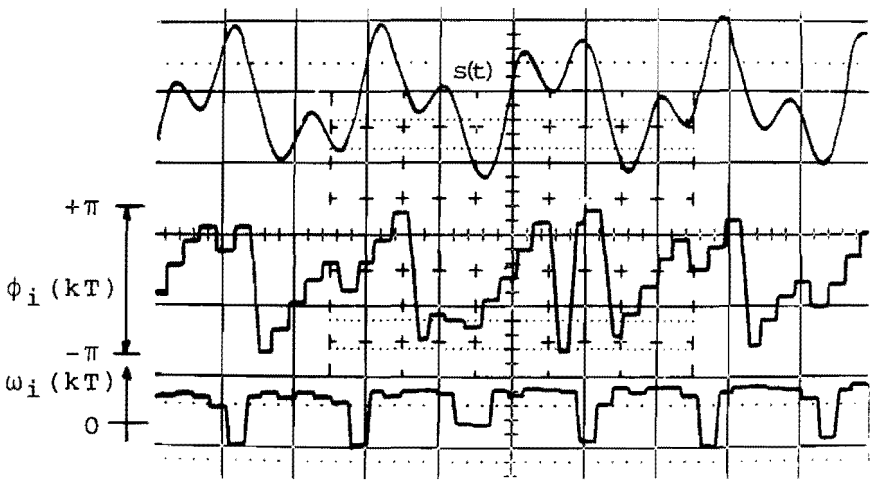


Fig. 3.21.3

Fig. 3.21 Sinusoid Plus Significant Second Harmonic

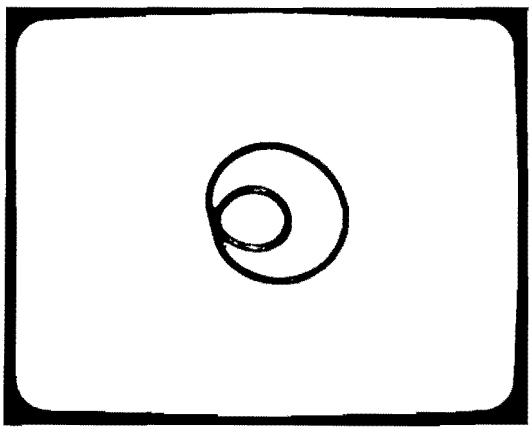


Fig 3.22.1

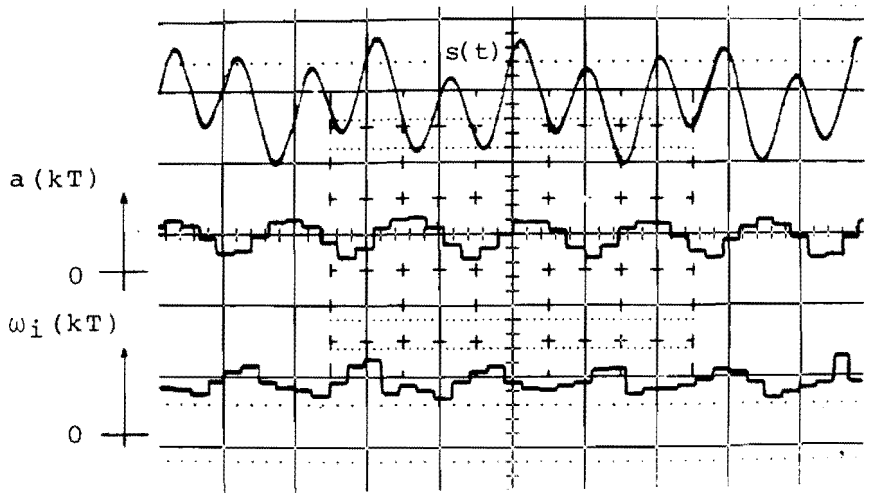


Fig. 3.22.2

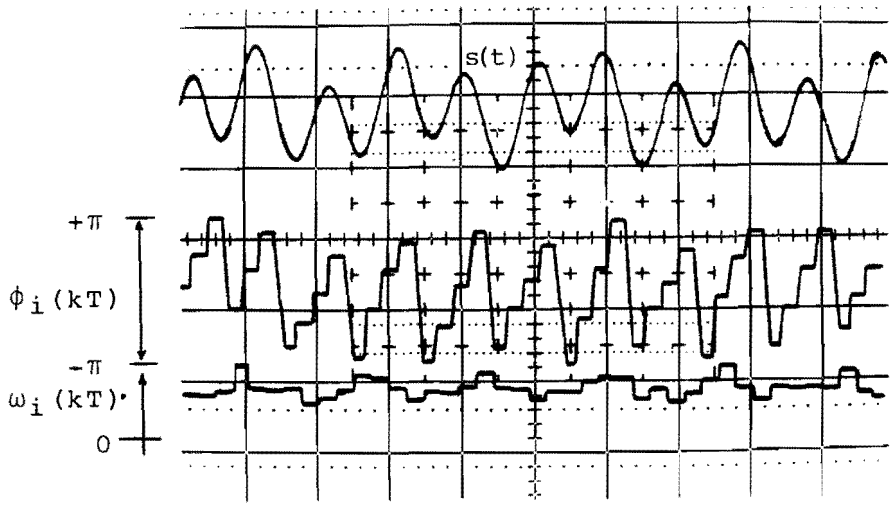


Fig. 3.22.3

Fig. 3.22 Sinusoid Plus Dominant Second Harmonic

This satisfied condition (c) and as expected, figures 3.22.2 and 3.22.3 display periodic instantaneous amplitude dips with corresponding instantaneous frequency rises and a doubling of the average phase slope.

Although the above tests indicated that the system was performing satisfactorily, the granular nature of the instantaneous frequency waveform cast doubt as its usefulness for the analysis of voiced speech sounds. For this reason, a further test was performed using a speech channel bandpass filtered sawtooth waveform to simulate a voiced sound. The sawtooth repetition rate was set at 100 Hz as this is around the fundamental frequency of a male speaker.

The vector plot of this periodic sound is shown in figure 3.23.1. There are four loops of the origin and one inner loop per cycle of the fundamental, indicating that the average rate of phase change is 800π radians per second and that there is one negative frequency excursion every 0.01 seconds.

Even though the instantaneous frequency plot of figure 3.23.2 is somewhat granular, the negative frequency excursion is quite evident, corresponding to a significant instantaneous amplitude dip. The instantaneous frequency waveform also exhibits a noticeable rise coinciding with a second significant instantaneous amplitude dip per cycle.

The phase plot of figure 3.23.3 verifies the predicted average rate of phase change by displaying four complete traversals of the range $-\pi$ to $+\pi$ per period of the fundamental. The occasional negative phase slopes correspond to negative instantaneous frequency excursions and other slight changes in the rate of phase change appear to correspond to shallow dips of the instantaneous amplitude waveform.

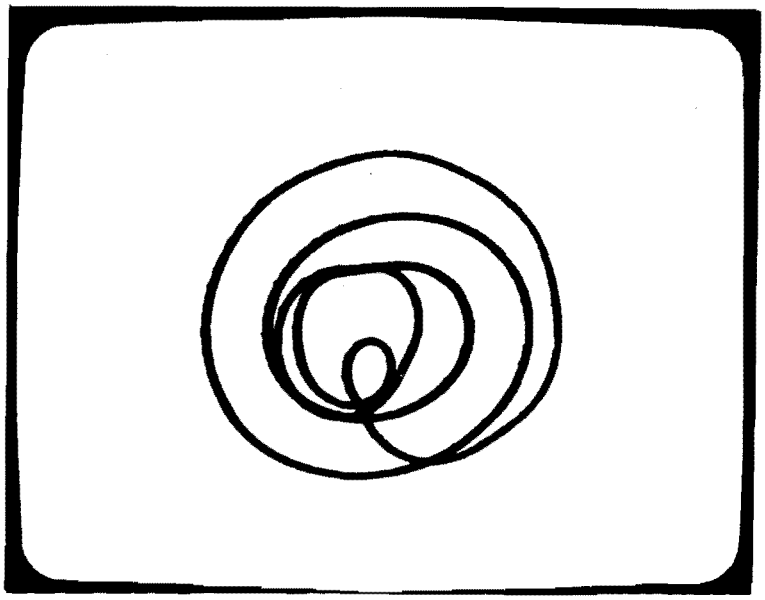


Fig. 3.23.1

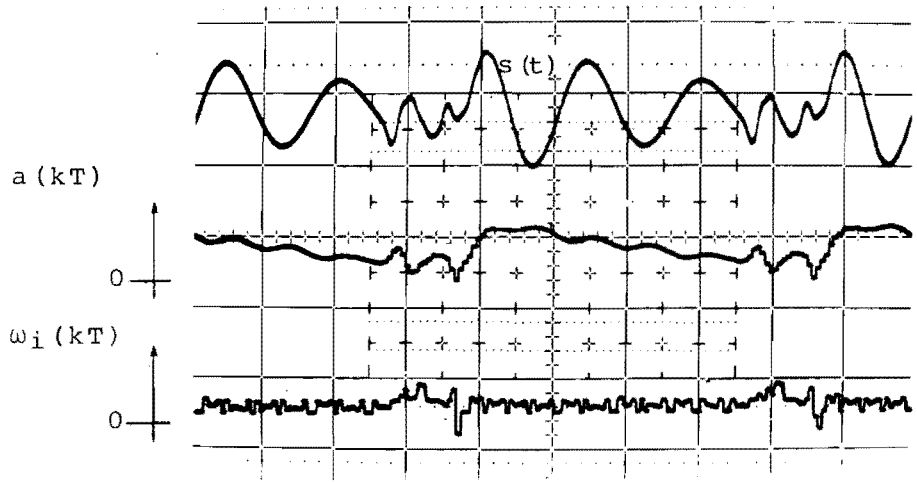


Fig. 3.23.2

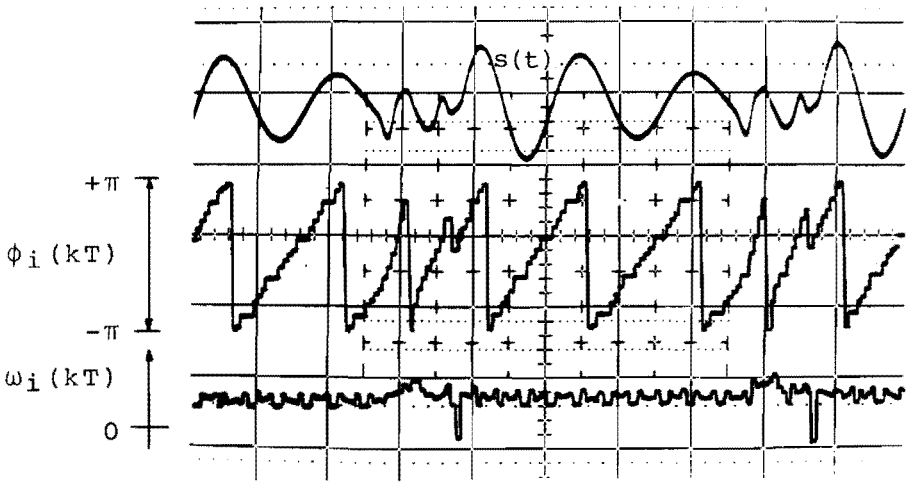


Fig. 3.23.3

Fig. 3.23 Sawtooth Test Signal Analysis

The relationship between the instantaneous waveforms and the location of analytic signal complex zeros developed in the examples of Chapter 2 suggests that the analytic waveform, corresponding to the real filtered sawtooth, has about 7 non-removed complex zeros per cycle. Two of these are significantly closer to the real axis than the others, one being LHP causing a negative instantaneous frequency excursion, and the other being UHP causing the noticeable instantaneous frequency rise.

Although the test signal of this example is significantly simpler than a real vowel, it demonstrates that the granular instantaneous frequency waveform, when viewed in conjunction with the vector plot and other waveforms, is sufficient to indicate major features.

(3.10) SPEECH ANALYSIS

Initial speech analysis was restricted to dealing with individual phonemes. Each phoneme yielded specific results in terms of its vector plot, instantaneous waveforms and averaged instantaneous waveforms, but the results for only two phonemes will serve to illustrate the type of output obtained.

Further speech analysis was in the form of the averaged instantaneous functions of words and phrases.

(3.10.1) VOICED PHONEME

A vector plot of the vowel /æ / uttered by a male speaker is illustrated in figure 3.24.1. Assuming a fundamental frequency of about 100 Hz, the display is over 3 cycles of the fundamental.

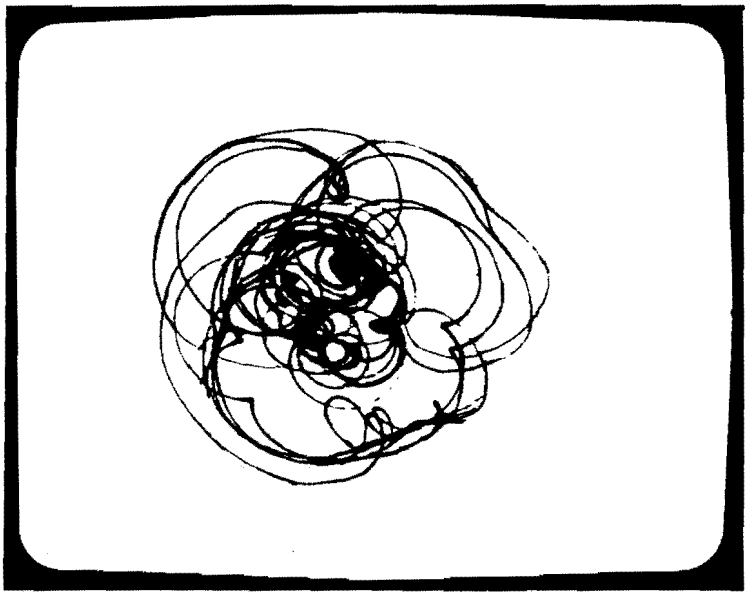


Fig. 3.24.1

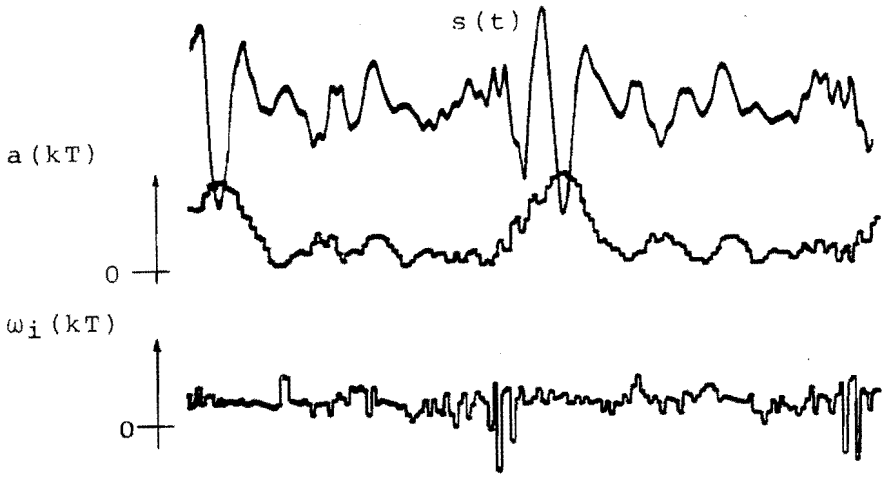


Fig. 3.24.2

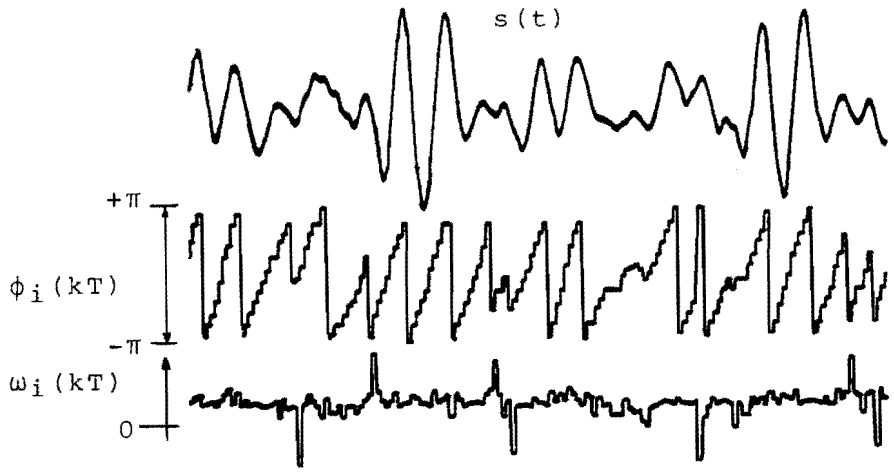


Fig. 3.24.3

Fig. 3.24 Waveforms for Vowel /æ/

Determining the average rate of phase change and number of negative instantaneous frequency excursions per cycle by the method of counting vector loops is not practical using this display. The untidy appearance is due to slight aperiodicities of the vowel waveform preventing the pattern from exactly overwriting itself, as in the example of the filtered sawtooth. The result is repeatable, however, as the general shape of the vector plot is reasonably stable in the long term.

This type of "roulette" figure representation of a phoneme, generated using narrow bandwidth phase splitting networks, has been used to help deaf children visualise the speech sounds they were trying to reproduce. (Ref. 106). In one such study, the vector loci were given the name "caligraphony" meaning "the beautiful writing of sounds". (Ref.107).

The displays of instantaneous waveforms, figures 3.24.2 and 3.24.3, reveal features that were masked by the unstable vector plot. Instantaneous amplitude exhibits 4 or 5 major dips per cycle and these correspond to disturbances of the instantaneous frequency waveform. During the period of high instantaneous amplitude, instantaneous frequency appears reasonably constant.

The instantaneous frequency waveform becomes negative at a rate of 2 or 3 times per cycle. Examination of the vector plot, however, leads us to expect a larger number of negative excursions and some are probably missed by the limited frequency response of the analyser.

The phase plot confirms the observed rate of negative excursions, and over one cycle it traverses the range $-\pi$ to π six times. The aperiodic nature of the vowel

is illustrated by differences between the reference waveform s(t) of figures 3.24.2 and 3.24.3.

Figures 3.25.1 and 3.25.2 are chart recordings of the instantaneous amplitude and frequency of the vowel /æ/ uttered out of context by the male speaker. The squelch option is being used in this case. The frequency response of the chart recorder, shown in figure 3.26, is such that most of the high frequency fluctuations of a(t) and $\omega_i(t)$ are removed.

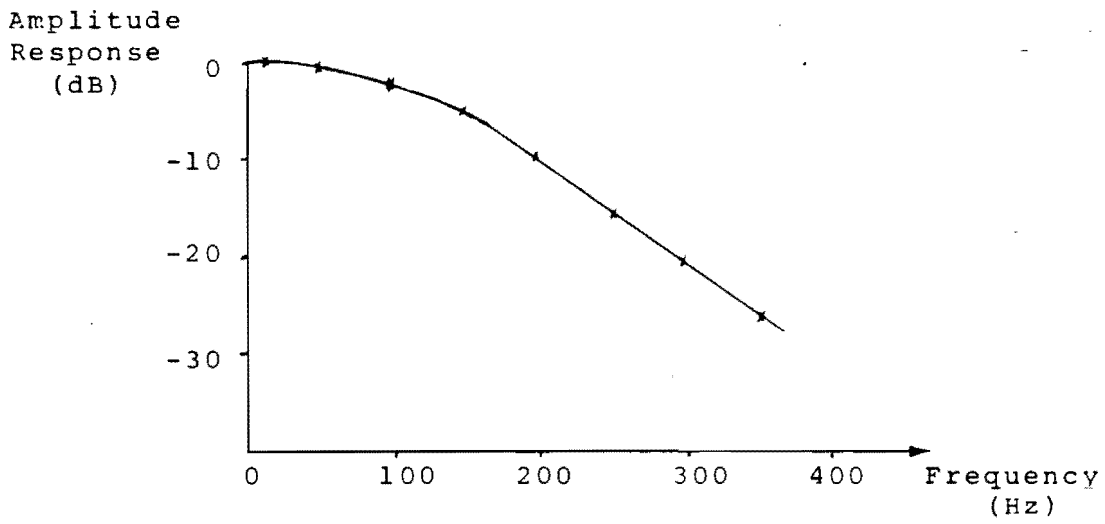


Fig. 3.26 Recorder Frequency Response

By the corresponding frequency scale, figure 3.25.2 shows that average instantaneous frequency of the vowel is between 500 Hz and 750 Hz. This is in agreement with figure 3.24.3 which shows the average to be 6 times the fundamental frequency.

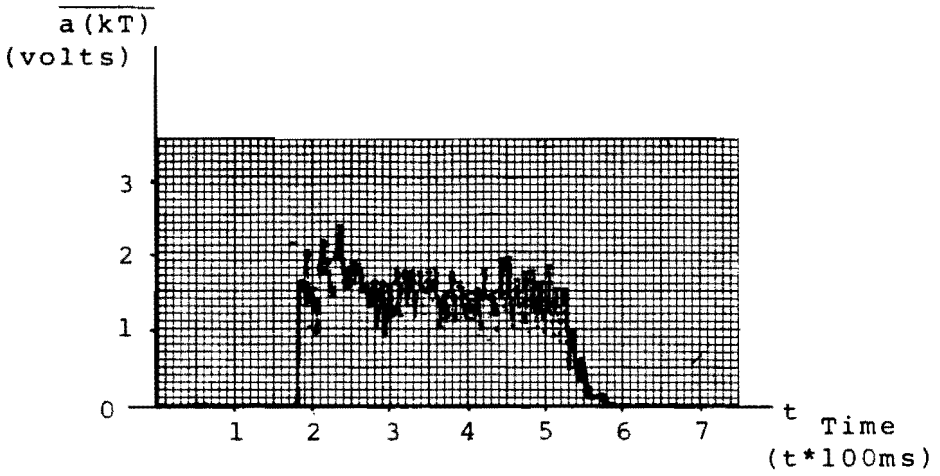


Fig. 3.25.1

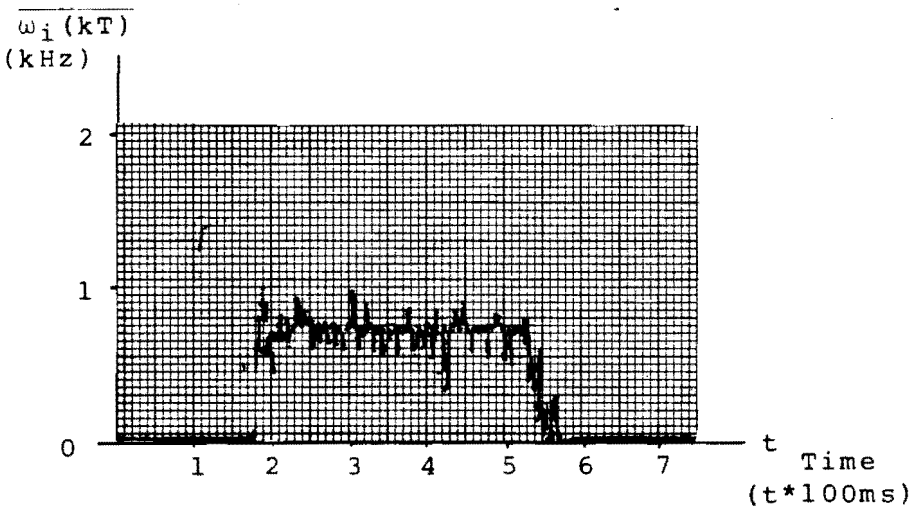


Fig. 3.25.2

Fig. 3.25 Time Averaged Instantaneous Waveforms of /æ/

Some early formant estimation schemes used zero crossing rate as an approximation to a vowels first formant frequency. (Ref. 1 and 81). Although not identical to zero crossing rate, the average instantaneous frequency of a vowel must also be a first formant frequency estimator and the approximate figure obtained above agrees well with 660 Hz, taken from Table 1.1,as the average first formant frequency of the vowel /æ /.

(3.10.2) UNVOICED PHONEME

Figure 3.27.1 is a vector plot of the unvoiced fricative /s/. The exposure time of this photograph is the same as for the voiced phoneme, figure 3.24.1. Resulting from a non-periodic noise like waveform, the vector plot appears untidy and estimation of the average rate of phase change by the method of counting loops is impossible.

The instantaneous waveforms of figures 3.27.2 and 3.27.3 are predictably non-periodic. As expected, however, dramatic dips or rises of instantaneous frequency still correspond to the significant dips of the instantaneous amplitude waveform.

Although the unvoiced fricative is not perfectly modelled by narrow band Gaussian noise, some similarities are evident between the instantaneous waveforms and the probability density functions outlined in Section 2.3.3.

It is difficult to say that the small section of instantaneous amplitude waveform shown in figure 3.27.2 fits a Rayleigh distribution, but the instantaneous phase waveform of figure 3.27.3 does appear to have samples evenly distributed between $-\pi$ and π. The instantaneous frequency samples of narrow band Gaussian noise should be symmetrically distributed around the band centre frequency, and this could be the case with the waveforms in figures 3.27.2 and 3.27.3.

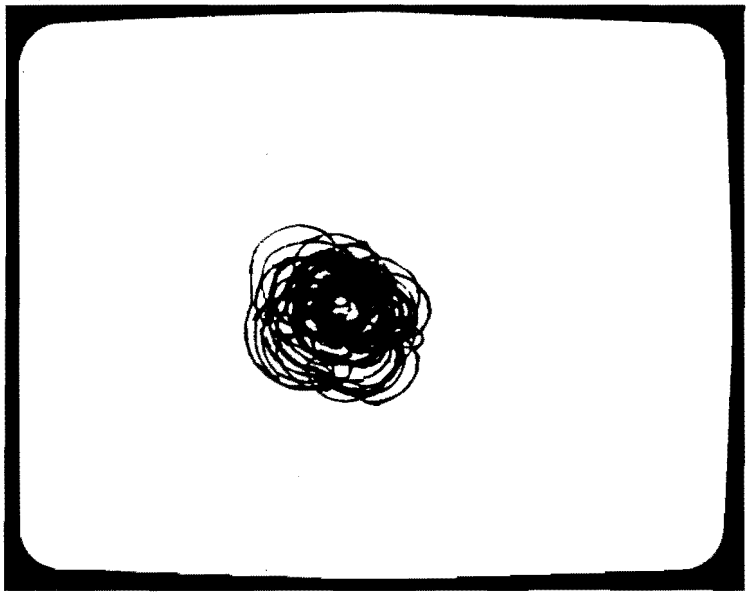


Fig. 3.27.1

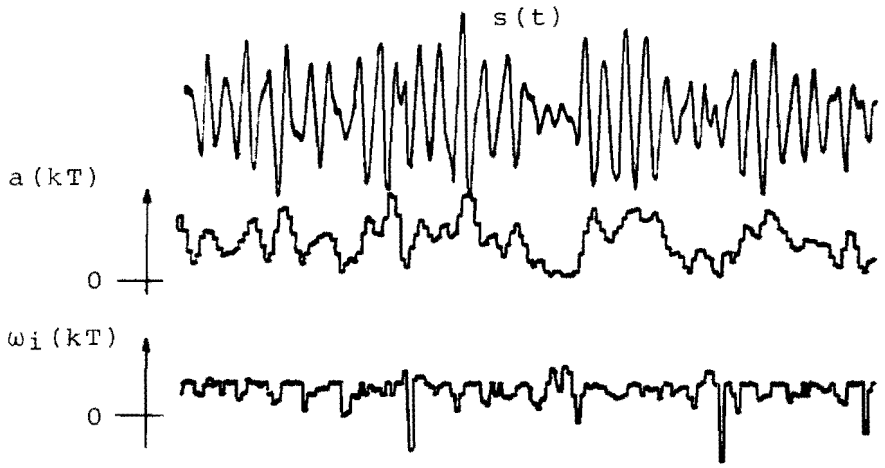


Fig. 3.27.2

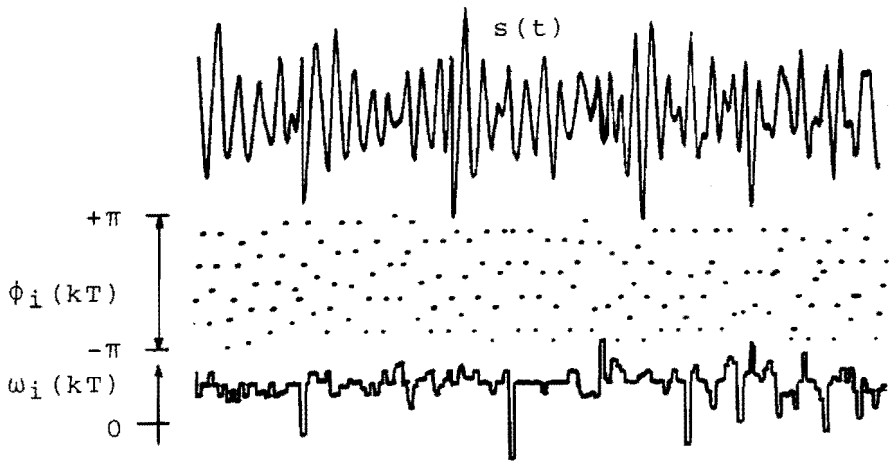


Fig. 3.27.3

Fig. 3.27 Waveforms for Unvoiced Fricative /s/

Instantaneous amplitude and frequency of the fricative /s/, uttered as part of the word "fox", are displayed on the chart recording figure 3.28. Unfortunately, the low pass characteristic of the chart recorder masks probability density function information which may otherwise have been evident from these recordings.

(3.10.3) PHRASE ANALYSIS

System output obtained during the anaylsis of whole words and phrases is illustrated by the chart recording figure 3.28. The phrase "The quick brown fox" was uttered by a male speaker.

The periods of zero instantaneous frequency between words or during inter-phoneme silences are artificially produced by the squelch logic. Although this makes the recordings easier to interpret, valuable instantaneous frequency information can be lost, such as at the beginning of the word "fox". The phoneme /f/ is of such low amplitude that the squelch logic is not properly triggered until the high amplitude /ɔ/.

Some information about changing average instantaneous frequency can be gleaned from the chart recording, but absence of the fine structure of instantaneous waveforms prevents a complete analysis.

All of the chart recordings presented above bear resemblance to plots of zero crossing rate for phonemes, words and phrases (Ref. 108 - 111). Although not identical, the recordings convey similar information.

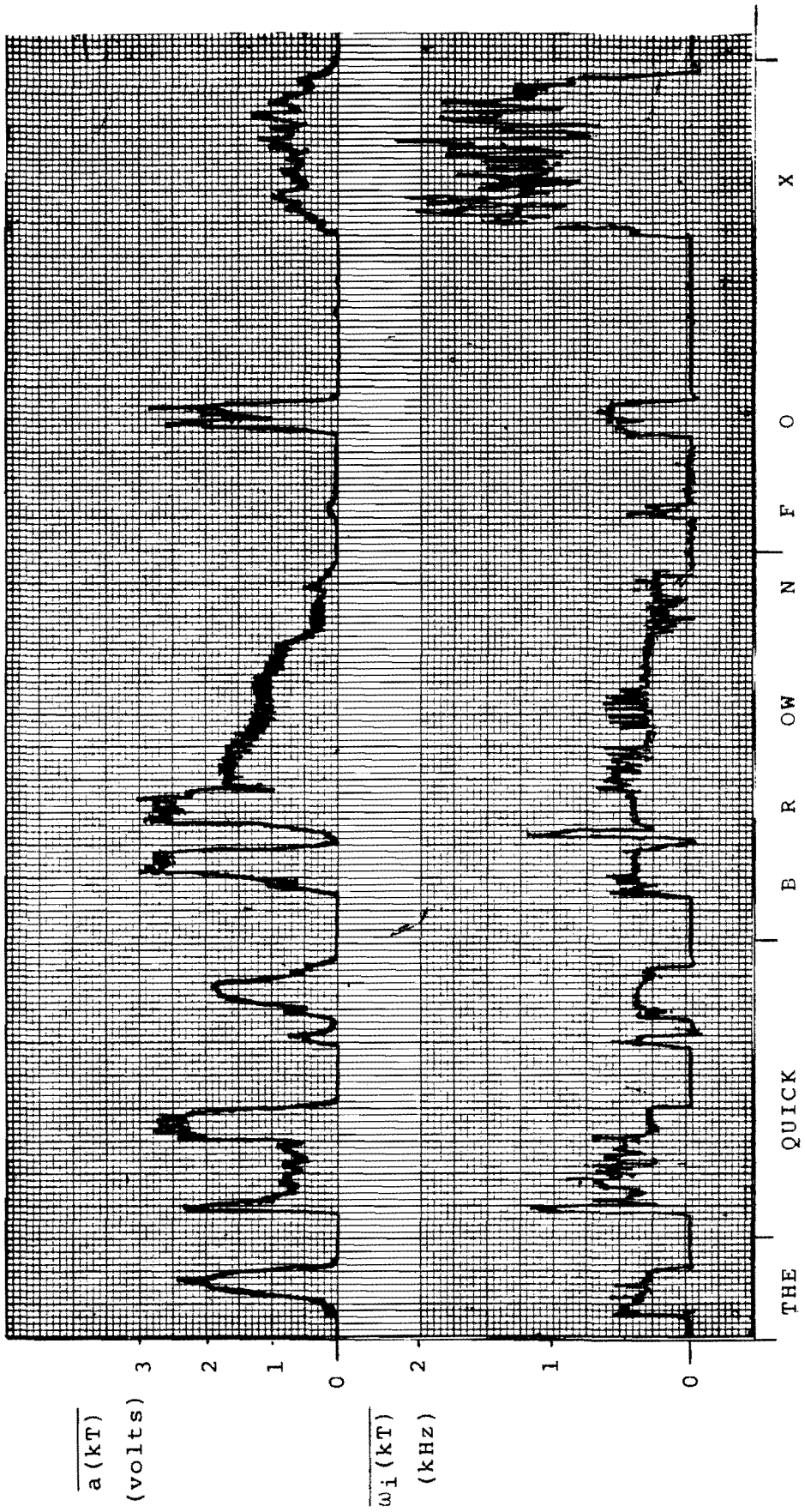


Fig. 3.28 Time Averaged Instantaneous Parameters of a Phrase

CHAPTER 4

(4.1) INTRODUCTION

Displays obtained using the hardware described in Chapter 3 suggested that further investigation of the instantaneous parameters of speech demands the generation of high fidelity instantaneous waveforms. Redesigning the analytic decoder to achieve this goal, however, requires increasing the sampling frequency and digital word length, with the corresponding increase in hardware complexity.

Fortunately, completion of analysis with the original hardware analytic decoder coincided with the introduction of A/D conversion facilities on the Electrical Engineering Departments VAX 11/780 digital computer. This facilitated the design of a computer based analytic decoder which, although incapable of real time operation, could perform 12 bit A/D conversions at rates of up to 80,000 per second, creating speech data records of several seconds length. The stored, sampled speech could then be digitally processed to produce and store the required instantaneous waveforms.

(4.2) DECODER DESIGN - SOFTWARE

Design and operation of the computer based analytic decoder parallels that of the hardware version.

Once a prefiltered speech waveform, s(t), is sampled and stored, it can be used to generate the orthogonal signal, $\hat{s}(kT)$. The instantaneous waveforms are then produced and stored on a sample by sample basis according to equations (4.1) and (4.2).

$$a(kT) = (s^2(kT)+\hat{s}^2(kT))^{\frac{1}{2}} \quad \ldots (4.1)$$

$$\omega_i(kT) = {}^{d}/_{dt}\{\tan^{-1}(\hat{s}(kT)/s(kT))\}. \ldots (4.2)$$

The digital computer is ideally suited to performing this type of calculation.

Vector loci are also generated on a sample by sample basis using stored instantaneous functions and the following relation.

$$\Psi(t)=a(t)\cos\{\int\omega_i(t).dt\}+j.a(t)\sin\{\int\omega_i(t).dt\} \quad . \; . \; . \quad (4.3)$$

The display of stored waveform segments and vector loci is by means of computer graphics terminals and hard copy plotters.

(4.2.1) HILBERT TRANSFORM

The use of a digital computer permits the implementation of a Hilbert Transform by means of Fourier Transform techniques or by simulation of an FIR filter. Having previously calculated a suitable impulse response, the FIR technique was chosen.

Unlike a TAD implementation, the computer based Gaussian windowed discrete Hilbert Transform impulse response is not limited to a fixed number of taps. Figures (4.1) and (4.2) illustrate the quadrature signal amplitude spectra for response lengths 15T and 203T ($T=1/f_s$ where f_s is the sampling frequency).

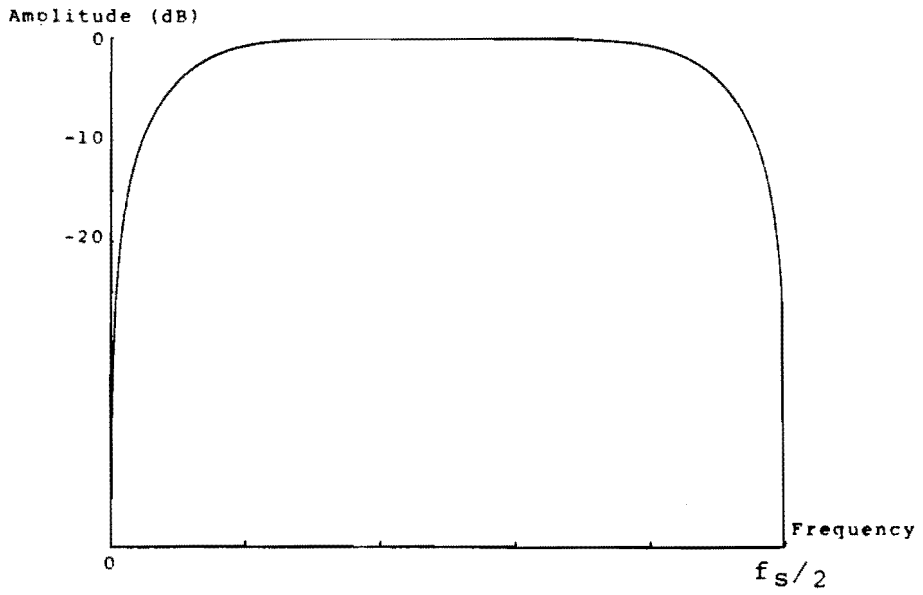


Fig 4.1 15 Tap Hilbert Transform Frequency Response

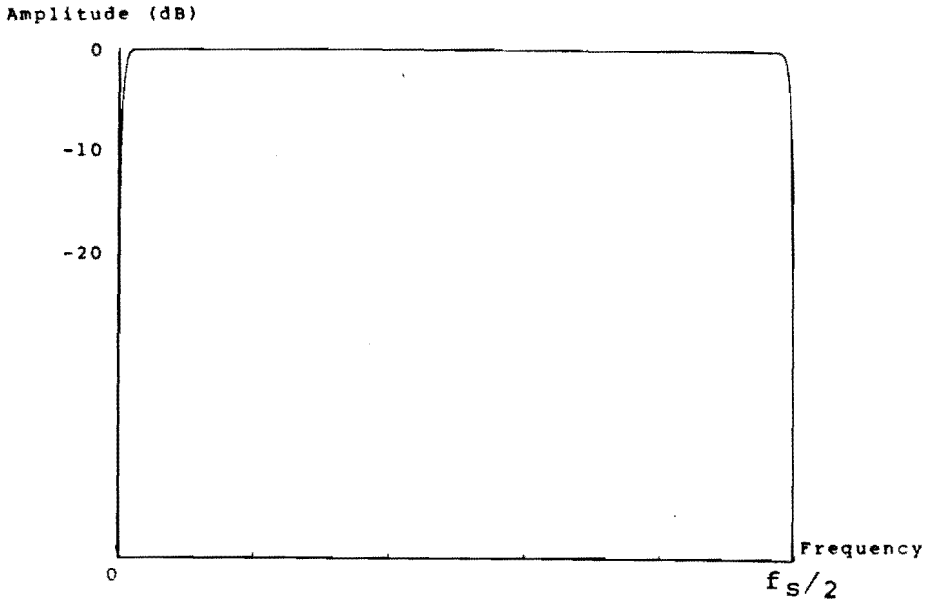


Fig 4.2 203 Tap Hilbert Transform Frequency Response

Although convolution with a long impulse response produces a better result in the frequency domain, it carries the penalties of requiring more processor time and of generating a quadrature signal with fewer usable samples. The process of losing samples through convolution is illustrated in figure (4.3).

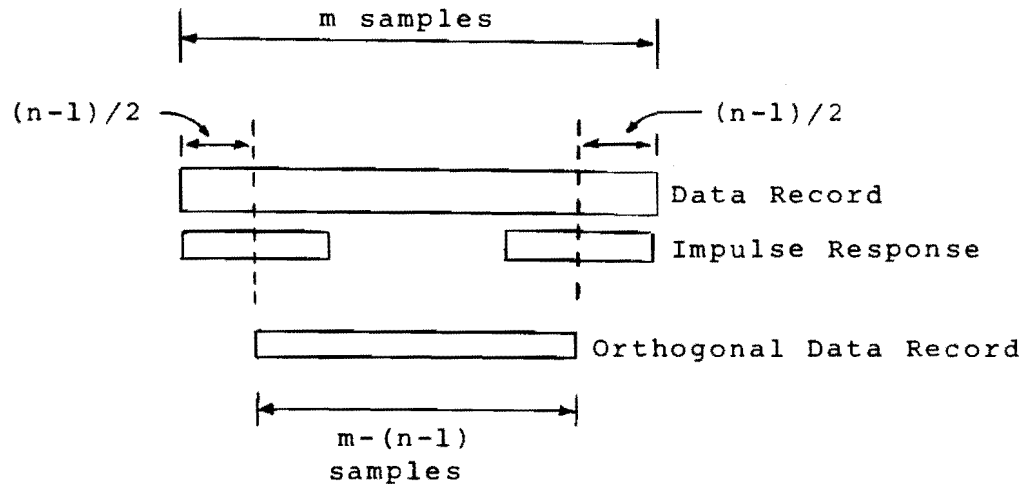


Fig 4.3 Generation of Orthogonal Signal

If the original data record is m samples long, convolution with an impulse response of n points results in an orthogonal signal of m-(n-1) useful samples. Orthogonal signal records are therefore packed with (n-1)/2 zeros at the beginning and end to ensure time correspondence of valid real and quadrature samples.

The use of a 203 point impulse response results in the loss of 202 usable samples, but this is considered negligible over a data record of many thousands of samples.

(4.3) SPEECH ANALYSIS

Satisfactory operation of the computer based system was verified using test signals previously applied to the hardware version. It was found that highly accurate results could be obtained if telephone bandwidth speech were sampled at 50,000 samples per second and $\hat{s}(kT)$ generated by a Hilbert transform impulse response length of 203 points.

As in the previous case, initial speech analysis was restricted to studies of individual phonemes. These are used to confirm and expand on earlier observations, leading finally to the analysis of words.

(4.3.1) VOWEL ANALYSIS

Analysis of the vowel /æ/, using the hardware analytic decoder, illustrated that the corresponding instantaneous waveforms and vector plots can be difficult to interpret. Computer analysis of a vowel confirms the intricate nature of the waveforms, but improved fidelity and control over the display of data facilitates analysis. The results of computer analysis of the vowel /ɛ/ are presented in figure (4.4).

Figure (4.4.1) is the time averaged amplitude spectrum of the vowel /ɛ/ uttered by a male speaker and prefiltered to the telephone bandwidth. This plot, and spectra presented later, was obtained by Fast Fourier Transform (FFT) techniques applied to 4096 points of vowel waveform data modified by a suitable windowing function.

The amplitude spectrum shows that the vowel is an almost periodic function of time with a fundamental frequency (line spacing) of around 90 Hz. Three resonances (first second and third formants) are visible.

Figures (4.4.2) and (4.4.3) are the instantaneous frequency and amplitude functions derived from two cycles (22 ms) of the vowels time waveform. As expected these waveforms are almost periodic and exhibit considerable fine structure. The large positive going spikes of instantaneous frequency (one per cycle) exceed the dynamic range of the plotting device and are clipped at the same level (+5,000 Hz).

The spectral composition of the instantaneous waveforms can be examined by suppressing their DC components and applying normal FFT analysis techniques. This results in the modified amplitude spectra of instantaneous frequency, figure (4.4.4), and instantaneous amplitude, figure (4.4.5).

Figure (4.4.4) reveals a strong spectral line structure at low frequency, which degenerates to a noise like spectrum at high frequency. The spectrum confirms the observation that instantaneous frequency is almost periodic and supports the prediction that instantaneous frequency is not bandlimited to the telephone bandwidth.

The spectrum, figure (4.4.5), shows that instantaneous amplitude is strongly periodic and apparently bandlimited. The frequency spacing of instantaneous amplitude resonances reflects the formant structure of the original vowel waveform.

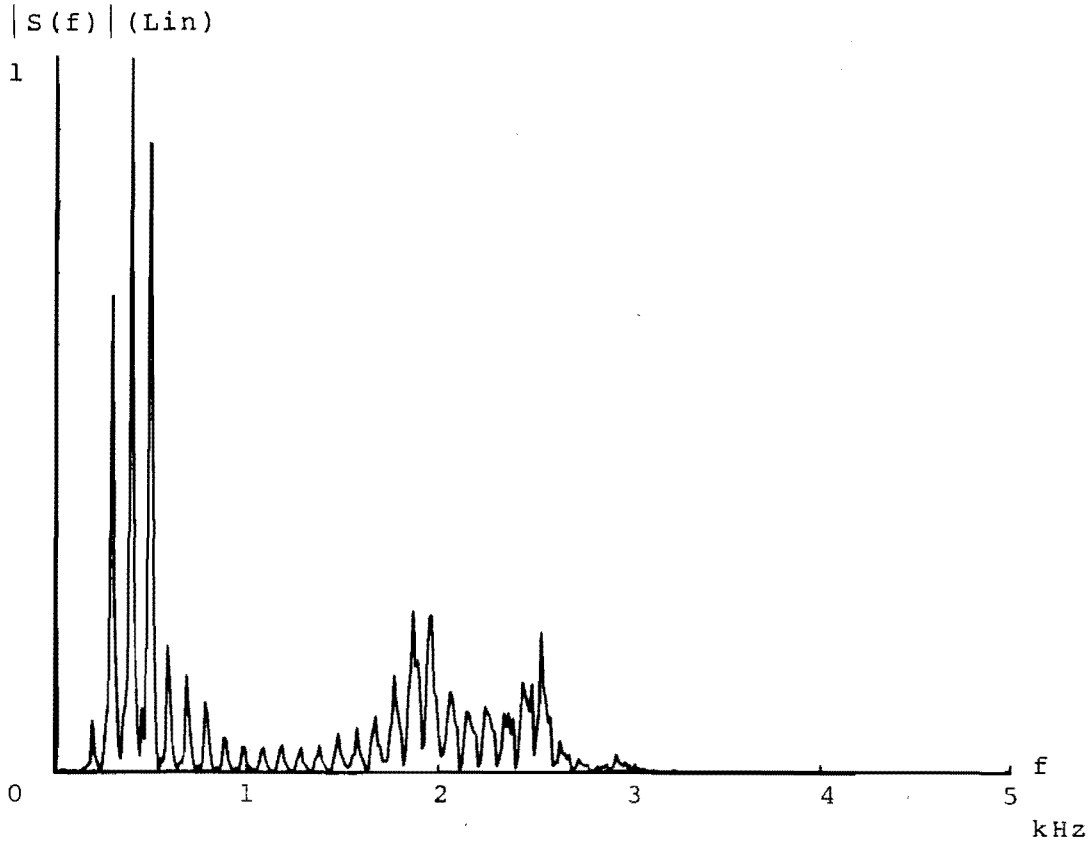


Fig 4.4.1

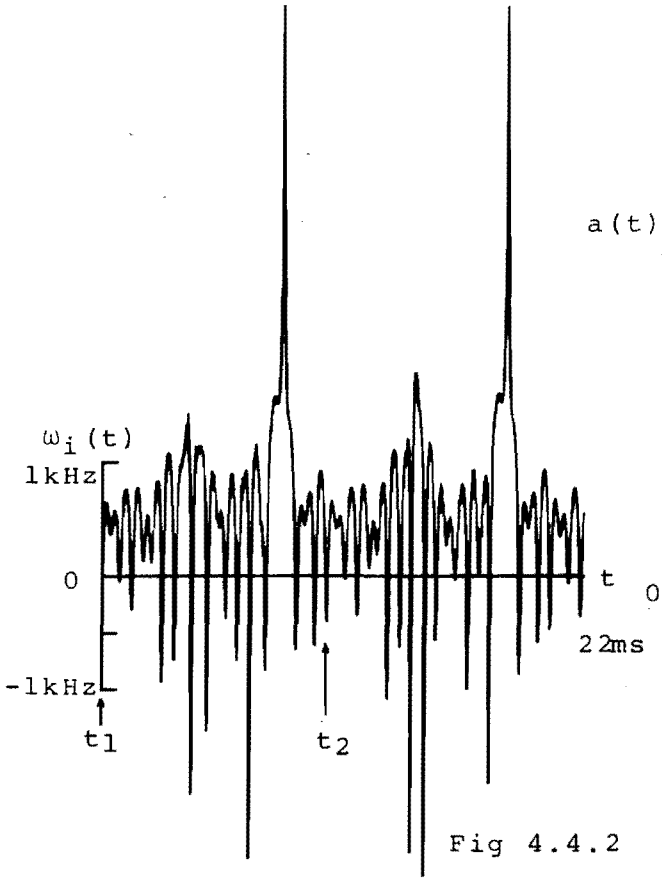


Fig 4.4.2

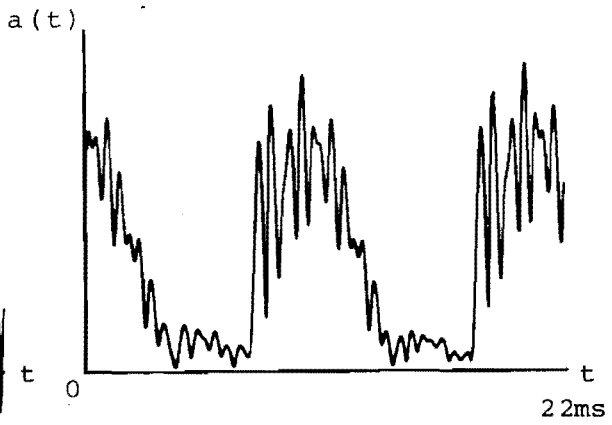


Fig 4.4.3

Figs 4.4.1-4.4.3 Analysis of /ε/

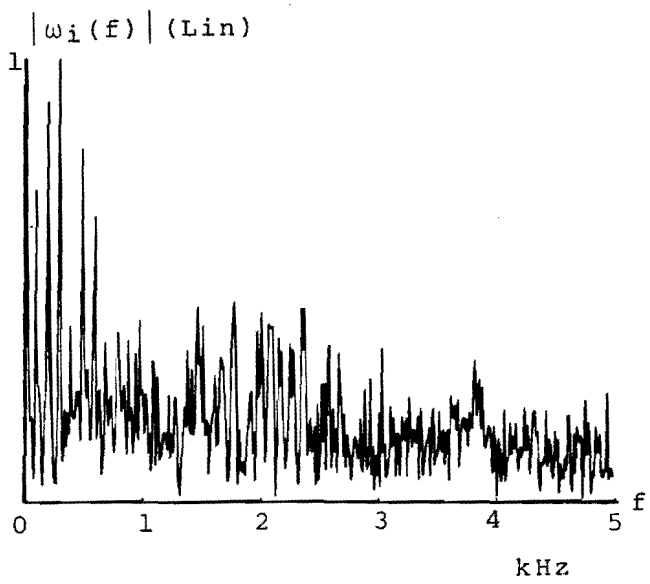


Fig 4.4.4

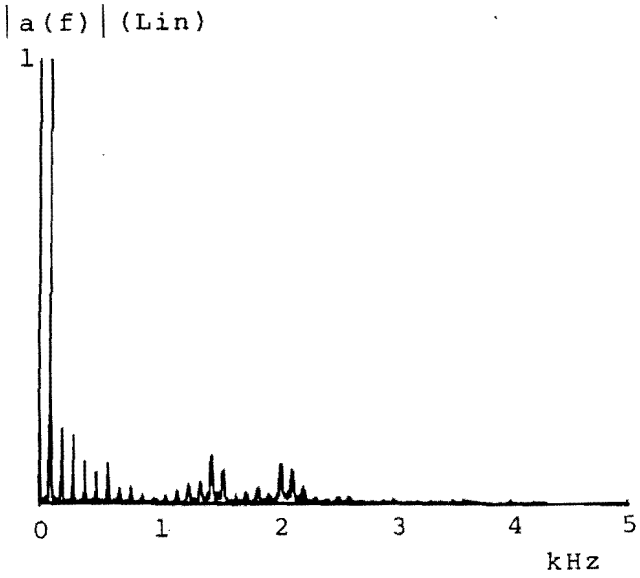


Fig 4.4.5

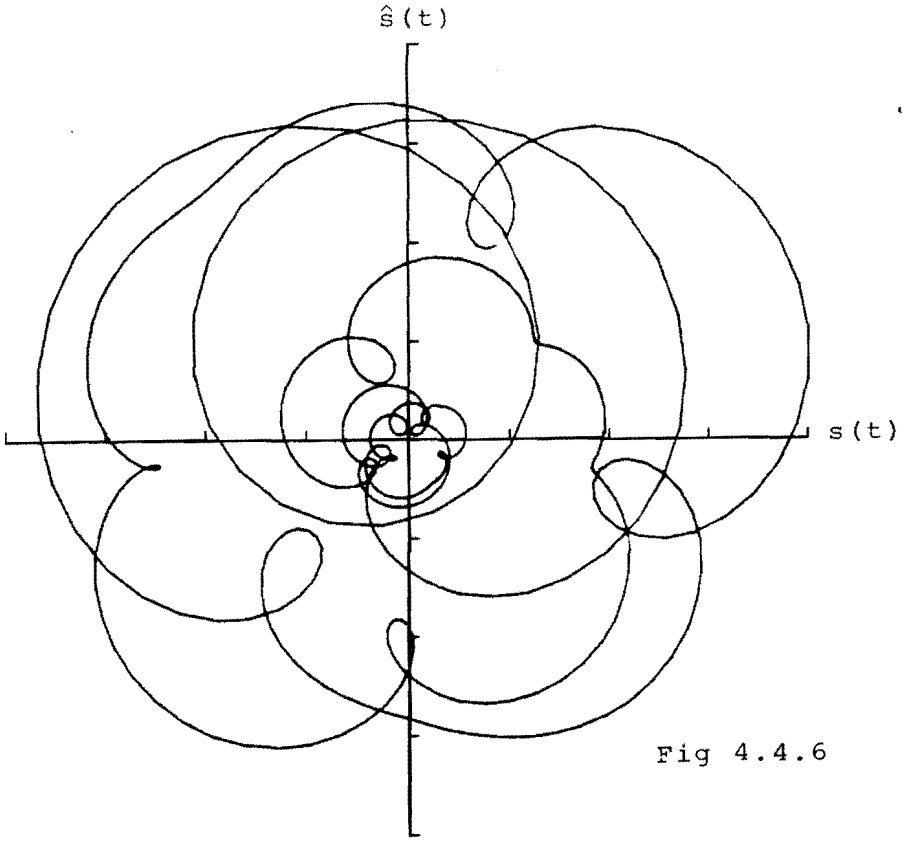


Fig 4.4.6

Figs 4.4.4-4.4.6 Analysis of /ɛ/

The vector diagram, figure (4.4.6), is plotted from instantaneous data over the period t_1 to t_2 indicated in figures (4.4.2) and (4.4.3). As only one cycle has been plotted, the locus is uncluttered and it is possible to count loops of the origin and inner loops. There are 6 full loops of the origin over one cycle of the fundamental frequency indicating that the average instantaneous frequency corresponds to the frequency of the 6th harmonic. The 13 inner loops correspond to 13 negative instantaneous frequency excursions per cycle, and this is confirmed by figure (4.4.2).

In accordance with the theory developed in Chapter 2, instantaneous amplitude dips correspond to instantaneous frequency rises or dips and these could be related to real waveform fluctuations and zero crossings. Of more interest in the case of vowels, however, is the relationship between instantaneous parameters and the information bearing characteristics of formant structure and spectral line spacing (pitch).

A strong relationship can be expected to exist between spectral structure and instantaneous parameters, as the positions and frequency of occurrence of analytic signal complex zeros are determinable from the Fourier series description of a periodic waveform. If a vowel is considered to be strictly periodic, with period T seconds, then the analytic signal corresponding to the real vowel waveform is represented by the Fourier series.

$$\Psi(z) = \sum_{m=n_1}^{n_2} A_m e^{jm\omega z} \qquad \ldots (4.4)$$

where An_1 ... An_2 are the complex Fourier coefficients, z is complex time

$$z = \tau + j\sigma \qquad \ldots (4.5)$$

and ω is the spectral line spacing, $\omega = 2\pi/T$. The analytic signal is restricted to the bandwidth $n.\omega$, where $n=n_2-n_1$.

Performing the substitution

$$\rho = e^{j\omega z} \qquad \ldots (4.6)$$

in equation (4.4) maps strips of the complex time plane, width T, onto the whole of the ρ plane and allows equation (4.4) to be rewritten as an algebraic polynomial in ρ

$$\Psi(z) = \sum_{m=n_1}^{n_2} A_m \rho^m \qquad \ldots (4.7)$$

As this is a polynomial of degree n, it possesses n zeros (roots) $\rho_1, \rho_2 \ldots \rho_n$, whose positions in the complex time plane are $z_1, z_2 \ldots z_n$ respectively. The zeros and their positions are related by the mapping equations

$$\rho_m = e^{j\omega z_m}, m = 1,2, \ldots n \qquad \ldots (4.8)$$

The zero positions $z_1, z_2, \ldots z_n$ and the constants A_{n_2} and n_1 fully define the periodic analytic signal $\Psi(z)$, by equation (4.7). (Ref. 92).

An important feature of this result is that any periodic analytic signal with a bandwidth of n spectral line spacings must exhibit n "non removed" complex zeros per cycle. The instantaneous amplitude and frequency waveforms corresponding to the analytic signal will therefore exhibit n fluctuations per cycle and these will vary in magnitude and polarity according to individual zero positions, (by equations (2.61) and (2.62)).

The relationship between analytic signal complex zeros and the number of spectral lines predicts that the instantaneous waveforms of figure (4.4) should exhibit around 26 spike or dip fluctuations per cycle. As this structure is not easily verified, low pass versions of /ɛ/ are

analysed separately.

(4.3.1.1) LOW PASS VOWEL ANALYSIS

1000Hz Low Pass

The first case, figure (4.5), is /ɛ/ low pass filtered to 1,000 Hz. The filtering is performed on the sampled data record by a software controlled linear phase FIR filter in such a way that the filtered waveform is not shifted in time with respect to the original data set. This allows the direct comparison of instantaneous waveforms and vector plots generated from different filtered data sets.

Figure (4.5.1) is the amplitude spectrum of the first formant of /ɛ/. The corresponding instantaneous waveforms, figures (4.5.2) and (4.5.3) have been considerably simplified, in comparison with the equivalent waveforms in figure (4.4). As the time interval is identical, these two sets of waveforms may be compared directly.

Interpretation of figures (4.5.2) and (4.5.3) suggests that the analytic signal has two major non removed UHP zeros per cycle. This agrees with theory as the first formant has an effective bandwidth of only 2 spectral line widths. The instantaneous frequency waveshapes, however, do not conform to the family of curves illustrated in Section (2.3.2.2) and the distortion is due to the influence of minor complex zeros, corresponding to the low amplitude spectral lines.

The vector plot, figure (4.5.4) shows 5 loops of the origin over one cycle of the fundamental, indicating that the average instantaneous frequency corresponds to the frequency of the 5th harmonic. It is notable that this is not the frequency of the dominant spectral component.

The vector plot also reveals a slight amplitude dip at time t_1 which, when viewed in conjunction with the corresponding low pass time waveform, figure (4.5.5), points to the presence of a minor LHP analytic signal complex zero.

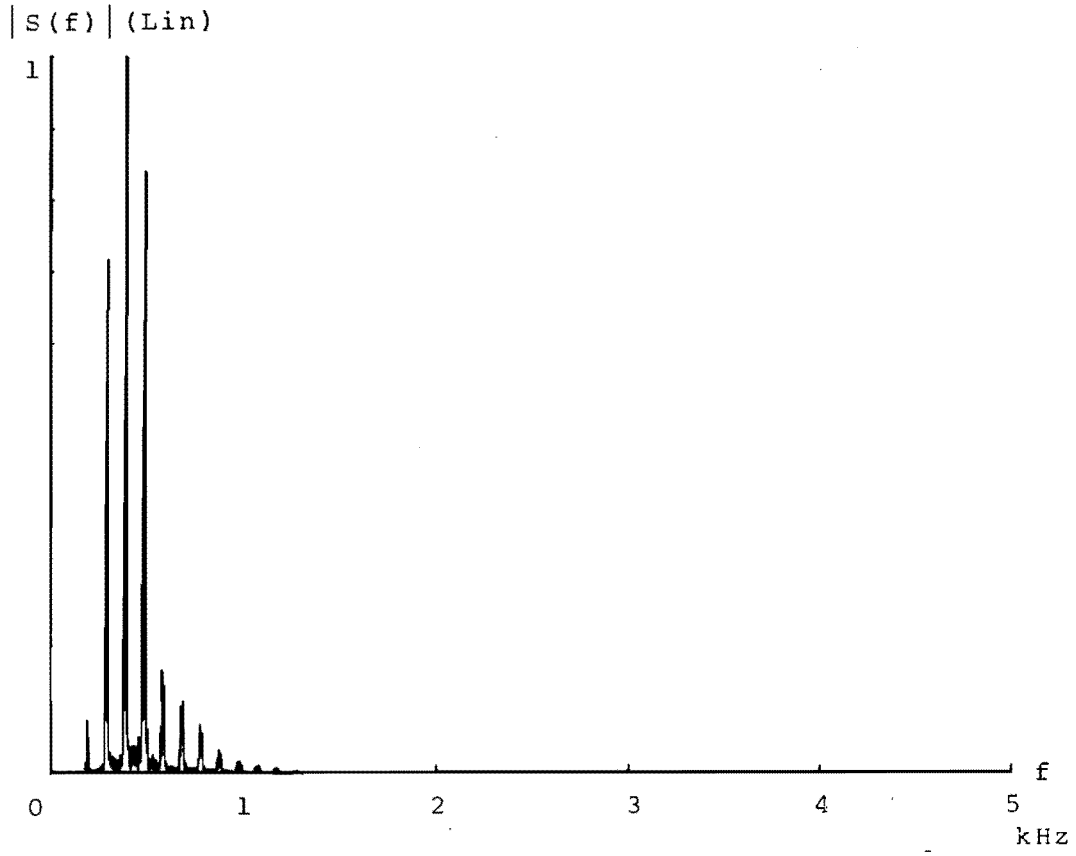


Fig 4.5.1

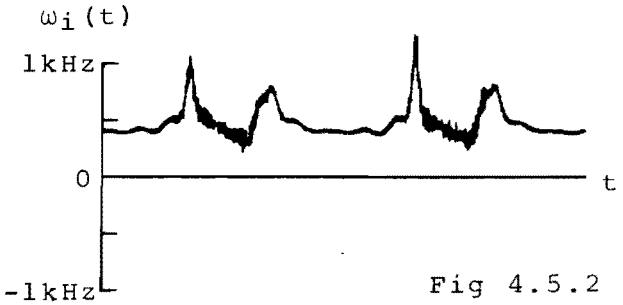


Fig 4.5.2

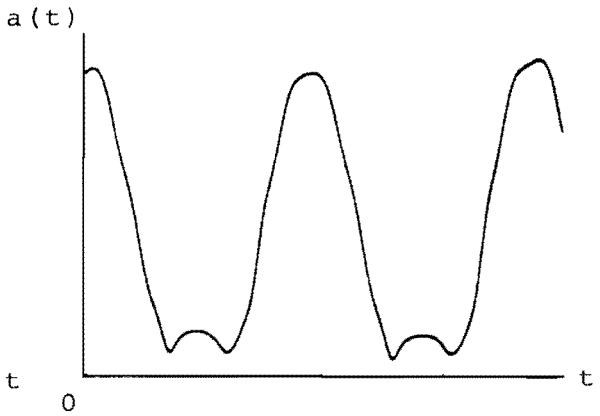


Fig 4.5.3

Figs 4.5.1-4.5.3 Analysis of Lowpass /ɛ/ (1000Hz)

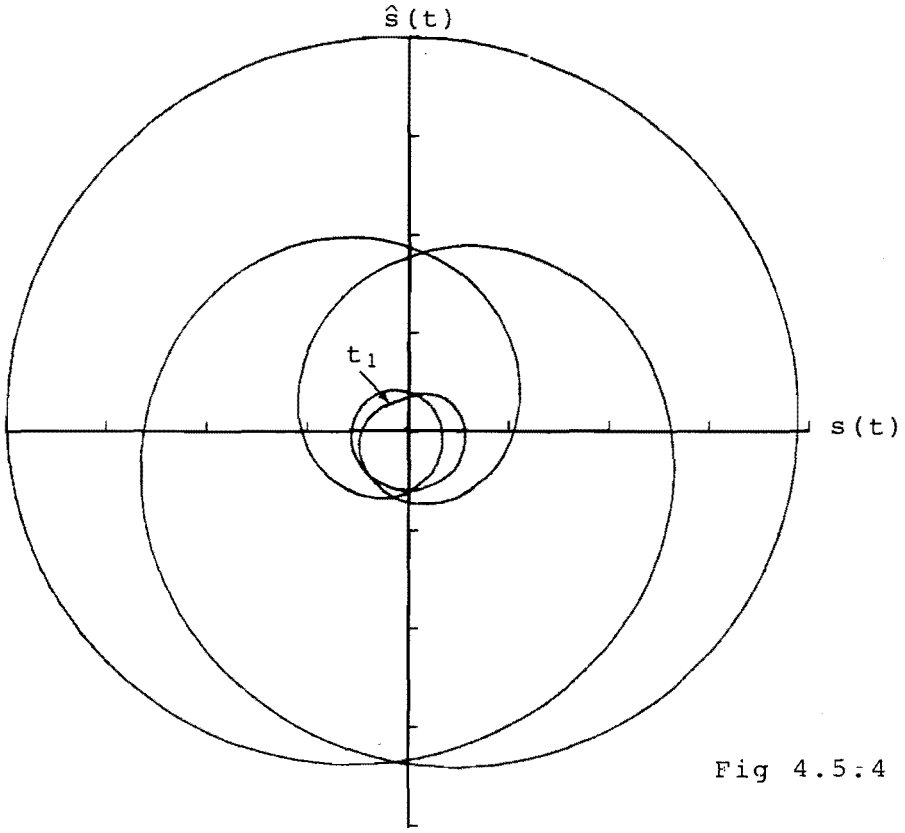


Fig 4.5.4

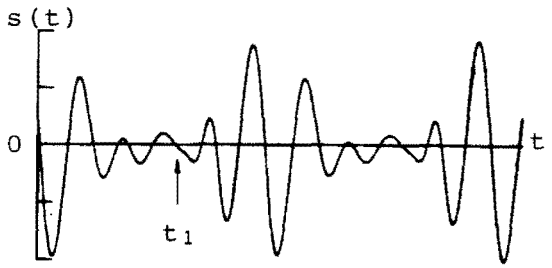


Fig 4.5.5

Figs 4.5.4-4.5.5 Analysis of Lowpass /ɛ/ (1000Hz)

In order to confirm the analysis of these waveforms it is convenient to construct a first formant model consisting of only three in-phase spectral lines, figure (4.6). For the first test, spectral amplitudes were set to A=2, B=3 and C=2.5.

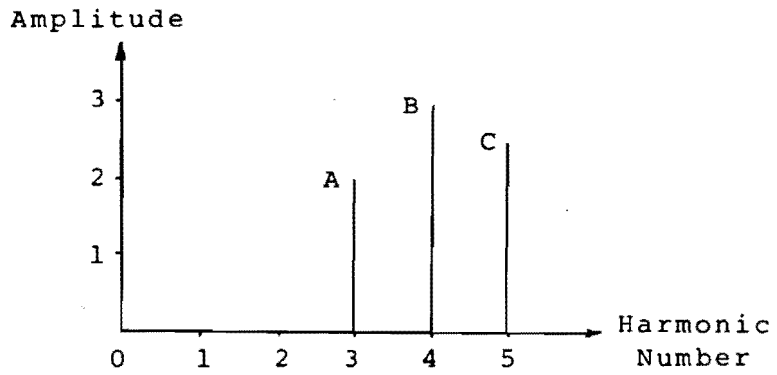


Fig. 4.6 First Formant Model

Figures (4.7.1) and (4.7.2) are the instantaneous waveforms derived from the model. The frequency curve now conforms to the expected family of shapes, indicating 2 UHP zeros per cycle. The average instantaneous frequency corresponds to the frequency of the 5th harmonic, and this is confirmed by the relevant vector plot, figure (4.8.1)

The series of vector plots, figure (4.8), show vector loci for the formant model with spectral amplitudes A=2, A=1 and A=0.5. It can be seen that for the case $A \leq 0.5$, one UHP zero per cycle becomes LHP and average instantaneous frequency drops to the frequency of the 4th harmonic (the dominant spectral component).

1600Hz Low Pass

The amplitude spectrum of /ɛ/ low pass filtered to 1,600 Hz, figure (4.9.1), shows that the signal now consists of the 3 major first formant components plus 15 upper harmonics.

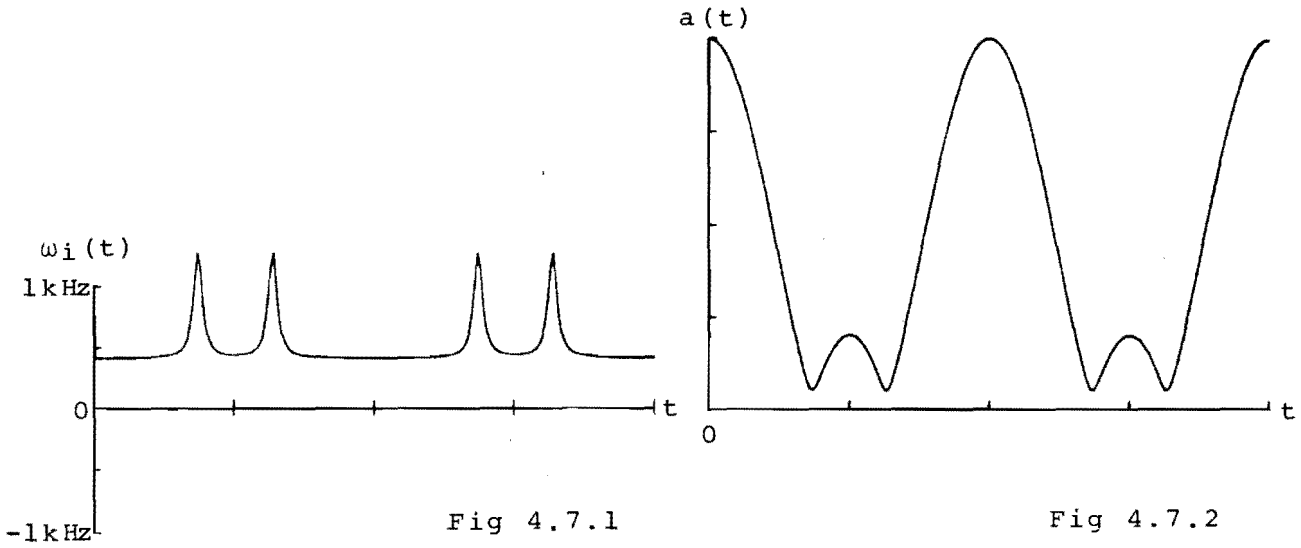


Fig 4.7.1 Fig 4.7.2

Fig 4.7 Instantaneous Parameters of First Formant Model

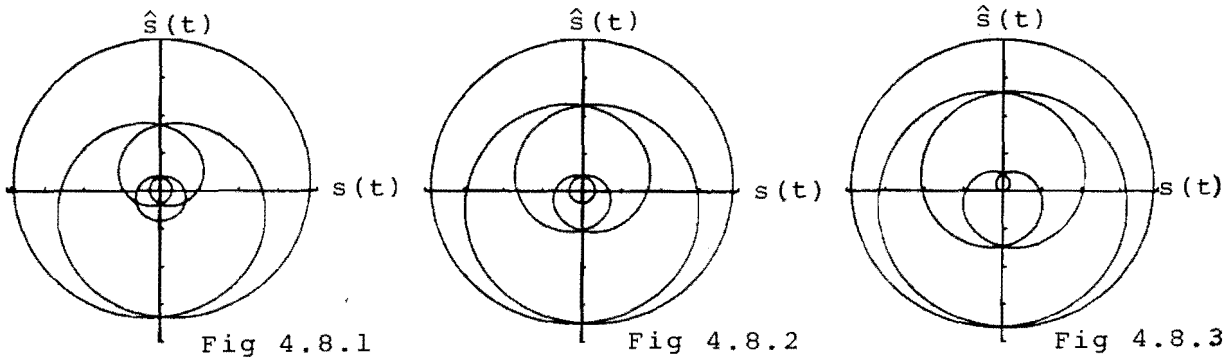


Fig 4.8.1 Fig 4.8.2 Fig 4.8.3

A=2 A=1 A=0.5

Fig 4.8 Vector Diagrams of First Formant Model

The instantaneous waveforms, figures (4.9.2) and (4.9.3), have altered significantly from the basic 2 UHP zeros per cycle shape. Both of the instantaneous waveforms now exhibit 14 or 15 almost evenly spaced dips per cycle. These are superimposed upon the 2 UHP zero per cycle waveshapes giving a total of around 17 non removed complex zeros per cycle. This agrees well with the signal bandwidth.

The vector plot, figure (4.9.4) confirms that the average instantaneous frequency still corresponds to the 5th harmonic and the vector now exhibits one inner loop per cycle.

2000Hz Low Pass

Low pass filtering /ɛ/ to 2,000 Hz, figure (4.10.1), allows the second formant to pass with almost no attenuation. The instantaneous waveforms, figures (4.10.2) and (4.10.3), now exhibit 15 evenly spaced and well defined dips per cycle, superimposed on the original 2 UHP zeros per cycle waveshapes. This is to be expected as the bandwidth of 15 line spacings between the 5th harmonic (corresponding to average instantaneous frequency) and the upper harmonic of the 2nd formant gives rise to 15 significant LHP complex zeros per cycle.

FFT analysis of the instantaneous frequency waveform (DC suppressed) confirms the presence of a spectral peak at about the frequency of the 15th harmonic, figure (4.10.4).

It is notable that LHP zeros occurring during periods of low average instantaneous amplitude cause sharp negative instantaneous frequency excursions (consistent with LHP analytic signal complex zeros near the real time axis). This effect is confirmed by the vector plot, figure (4.10.5), which shows vector inner loops occurring close to the origin and giving rise to high negative angular velocities.

The periods of low average instantaneous amplitude (one per cycle) are due to the proximity in real time of the two major UHP zeros associated with the first formant, or equivalently, are due to the almost in phase spectral components of the first formant adding destructively over part of a period.

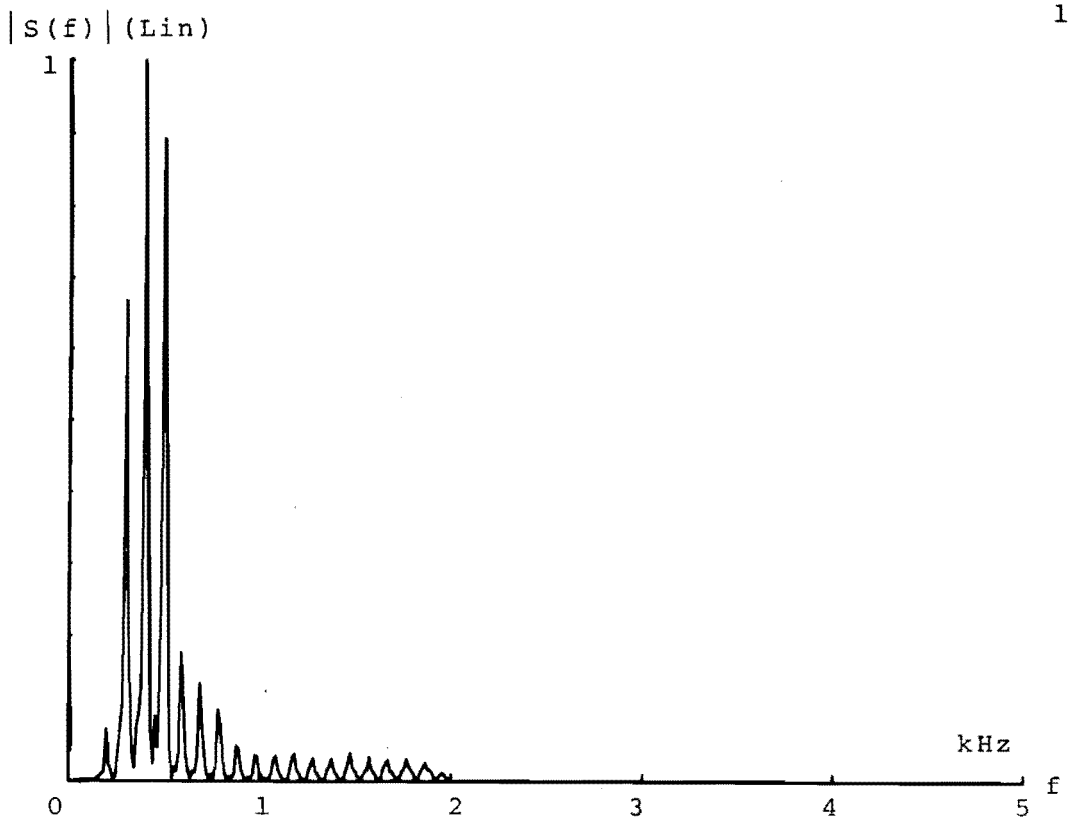


Fig 4.9.1

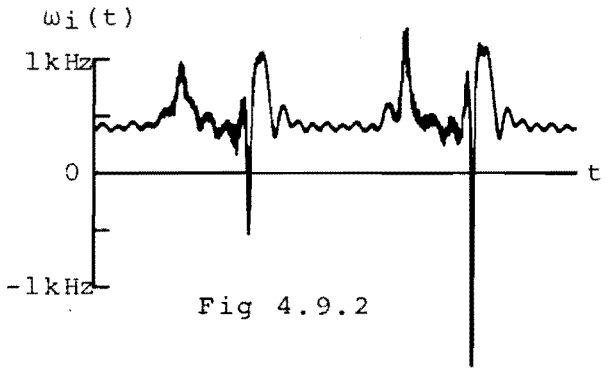


Fig 4.9.2

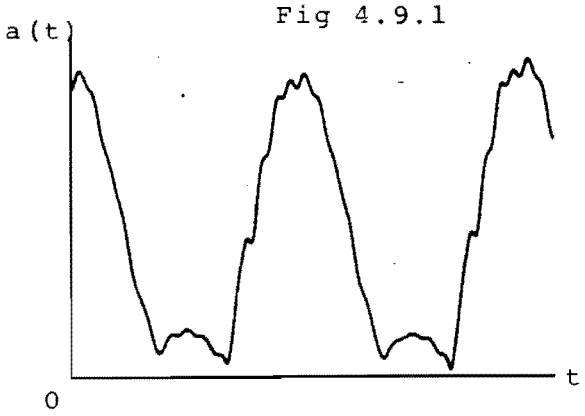


Fig 4.9.3

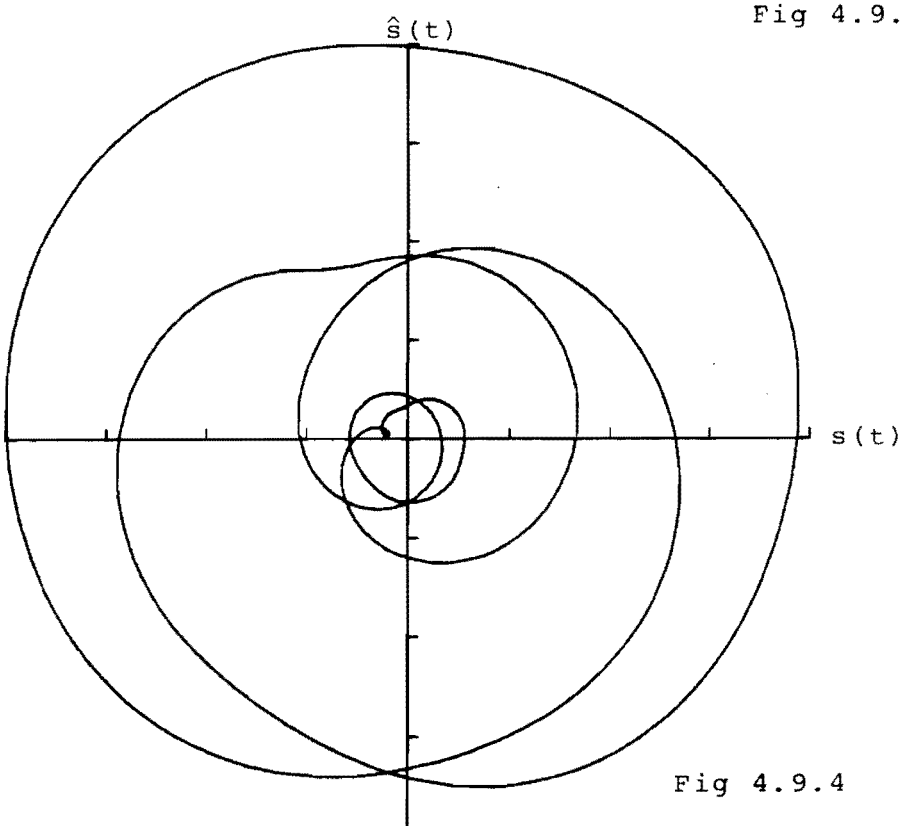


Fig 4.9.4

Fig 4.9 Analysis of Lowpass /ε/ (1600Hz)

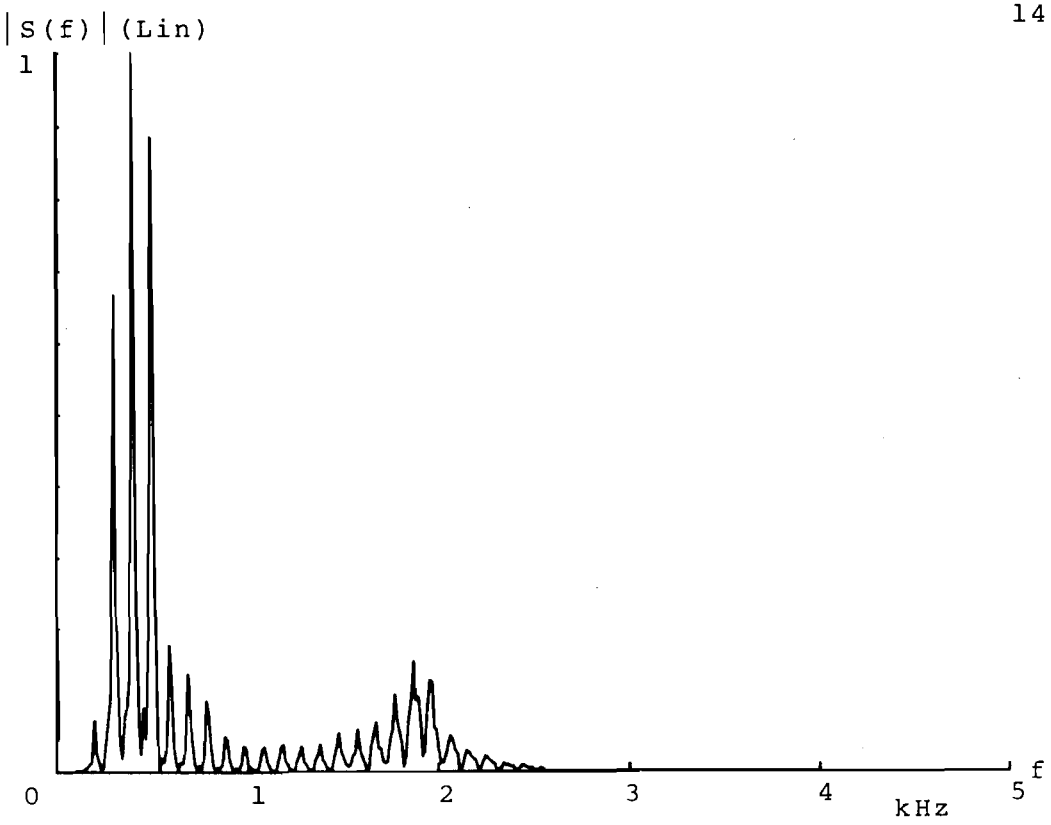


Fig 4.10.1

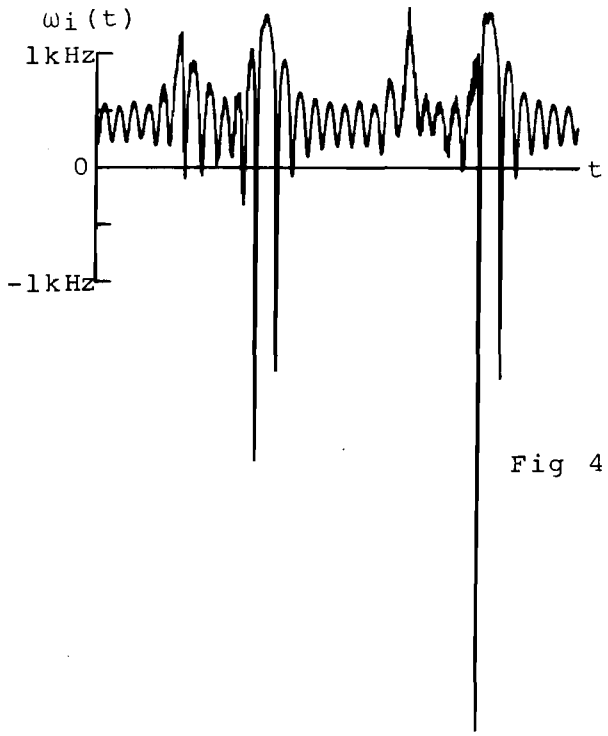


Fig 4.10.2

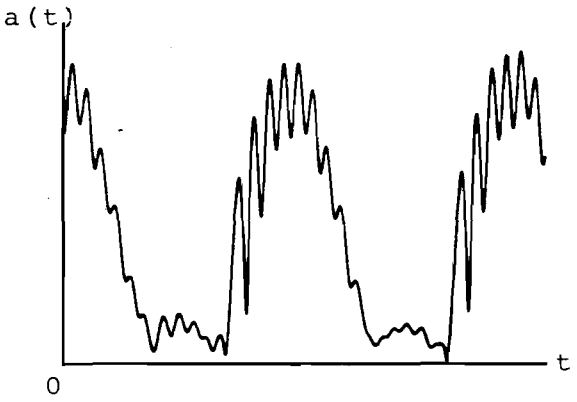


Fig 4.10.3

Figs 4.10.1-4.10.3 Analysis of Lowpass /ɛ/ (2000Hz)

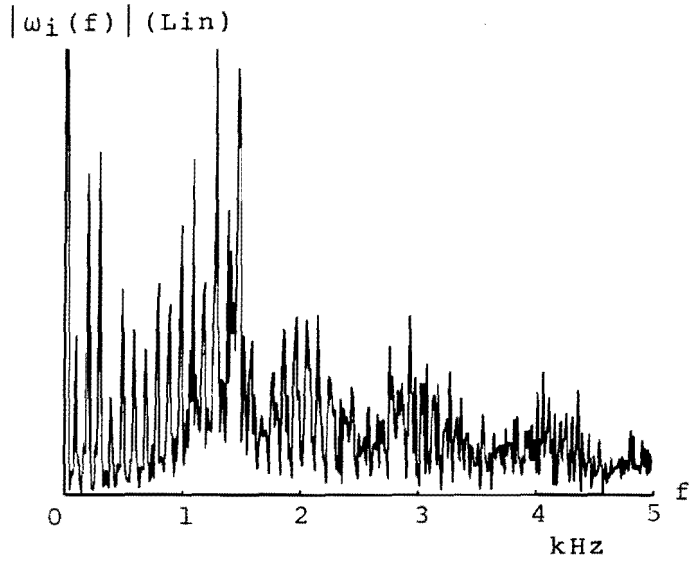


Fig 4.10.4

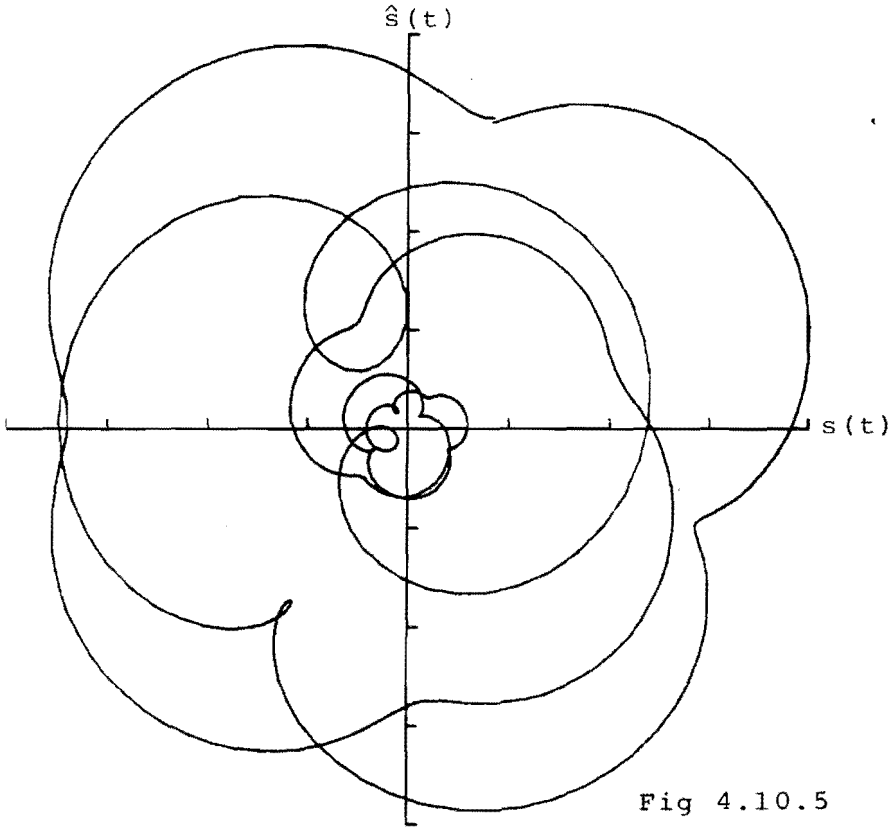


Fig 4.10.5

Figs 4.10.4-4.10.5 Analysis of Lowpass /ɛ/ (2000Hz)

2200Hz Low Pass

Increasing the low pass cutoff to 2,200 Hz passes the full second formant plus some higher harmonics, figure (4.11.1). The instantaneous waveforms, figures (4.11.2) and (4.11.3), still display about 15 dips per cycle superimposed on the 2 UHP zero waveshapes, but the instantaneous amplitude dips are of greater magnitude and the instantaneous frequency dips more irregular than in the previous case. Instantaneous frequency dips exceed the plotters range, at -6,000 Hz, once per cycle and these correspond to the large vector inner loop which passes close to the origin, figure (4.11.4).

2400Hz Low Pass

The final low pass version is /ɛ/ filtered to 2,400 Hz. In this case, all formants are present, although the third is of reduced amplitude, figure (4.12.1). Inspection of the low pass vowel spectrum leads to the prediction that there will now be 21 superimposed instantaneous parameter dips per cycle, corresponding to the 21 spectral lines above the 5th harmonic. Examining the instantaneous waveforms, figures (4.12.2) and (4.12.3), and vector plot, figure (4.12.5), however reveals that this is not the case.

With the increase in magnitude of upper harmonics, the large vector inner loop of figure (4.11.4) has expanded to encompass the origin, figure (4.12.5). This has led to the conversion of one LHP zero to UHP and has correspondingly increased the average instantaneous frequency to that of the 6th harmonic. The new UHP zero causes instantaneous frequency to clip at +5,000 Hz once per cycle, figure (4.12.2).

Taking the zero conversion into account, instantaneous frequency should now exhibit one new spike per cycle, and 20 dips, superimposed on 2 rises associated with the UHP zeros of the first formant. There are now about 18 countable dips per cycle on figure (4.12.2). The spectral analysis of instantaneous frequency, figure (4.12.4) shows no significant spectral peak at the 20th harmonic, but energy spread between the 14th and 21st.

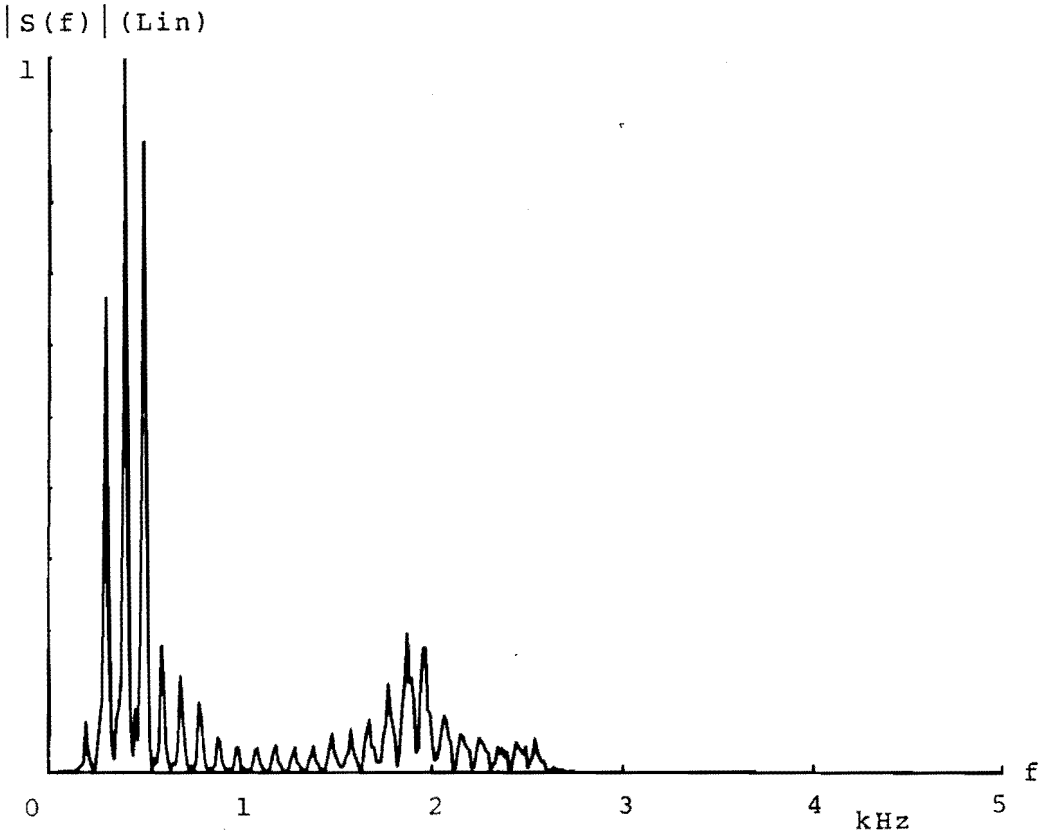


Fig 4.11.1

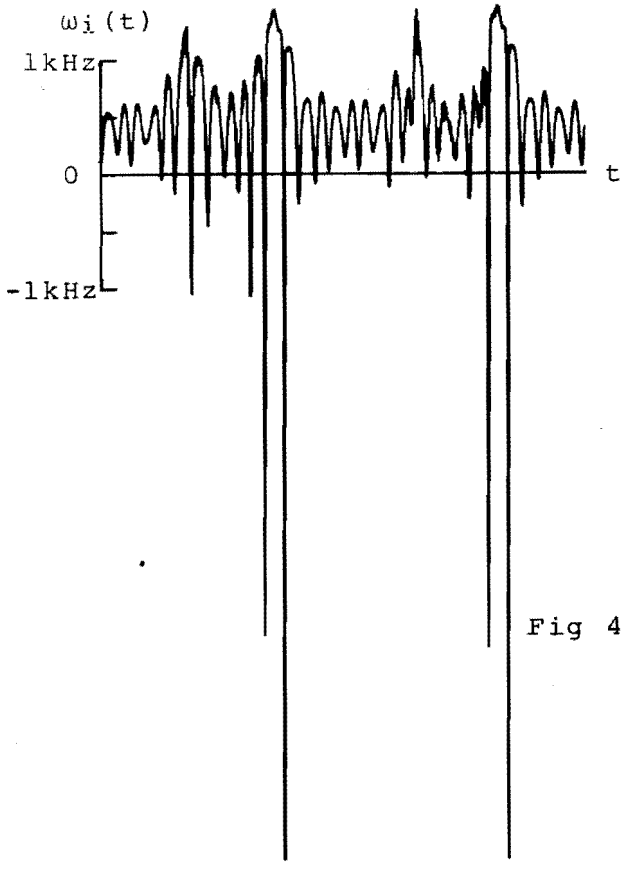


Fig 4.11.2

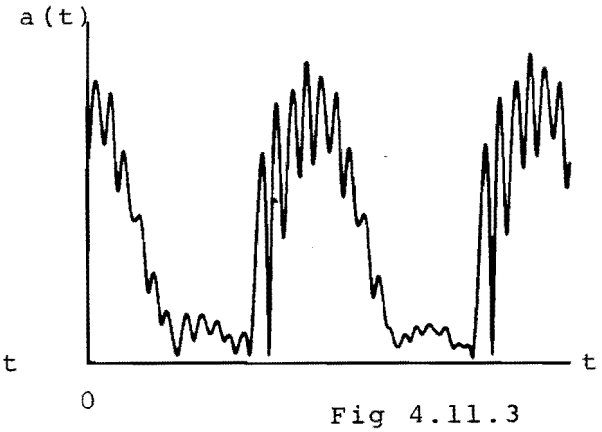


Fig 4.11.3

Figs 4.11.1-4.11.3 Analysis of Lowpass /ɛ/ (2200Hz)

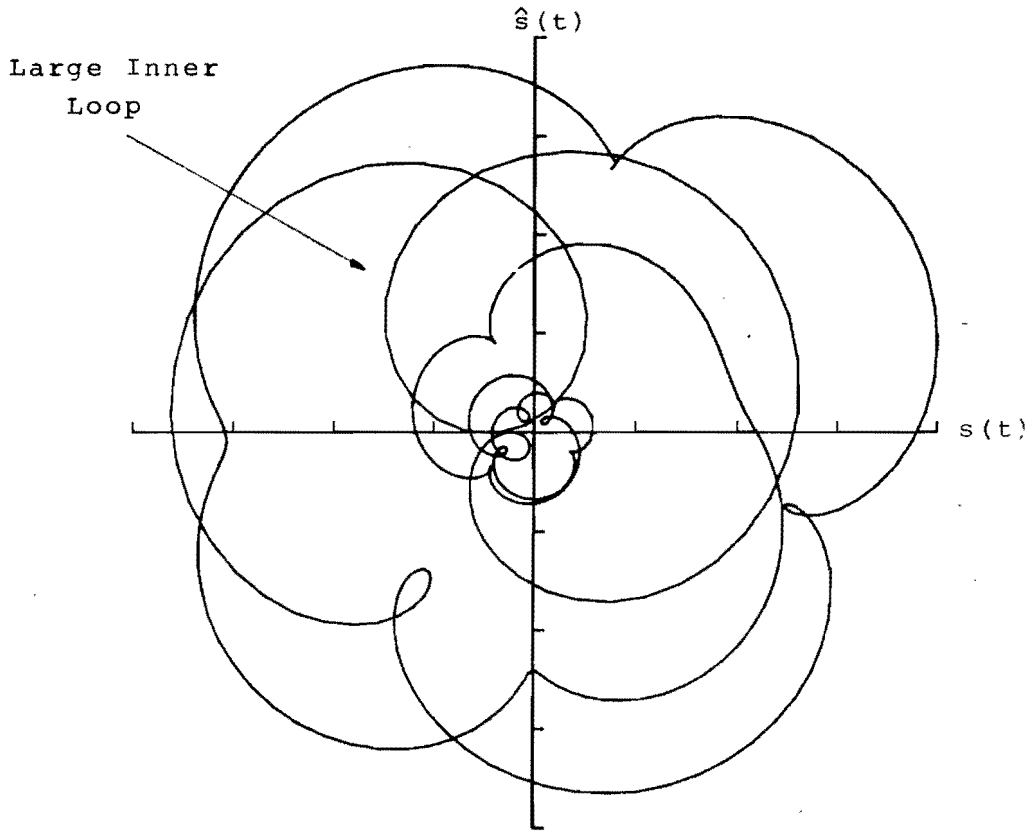


Fig 4.11.4 Analysis of Lowpass /ɛ/ (2200Hz)

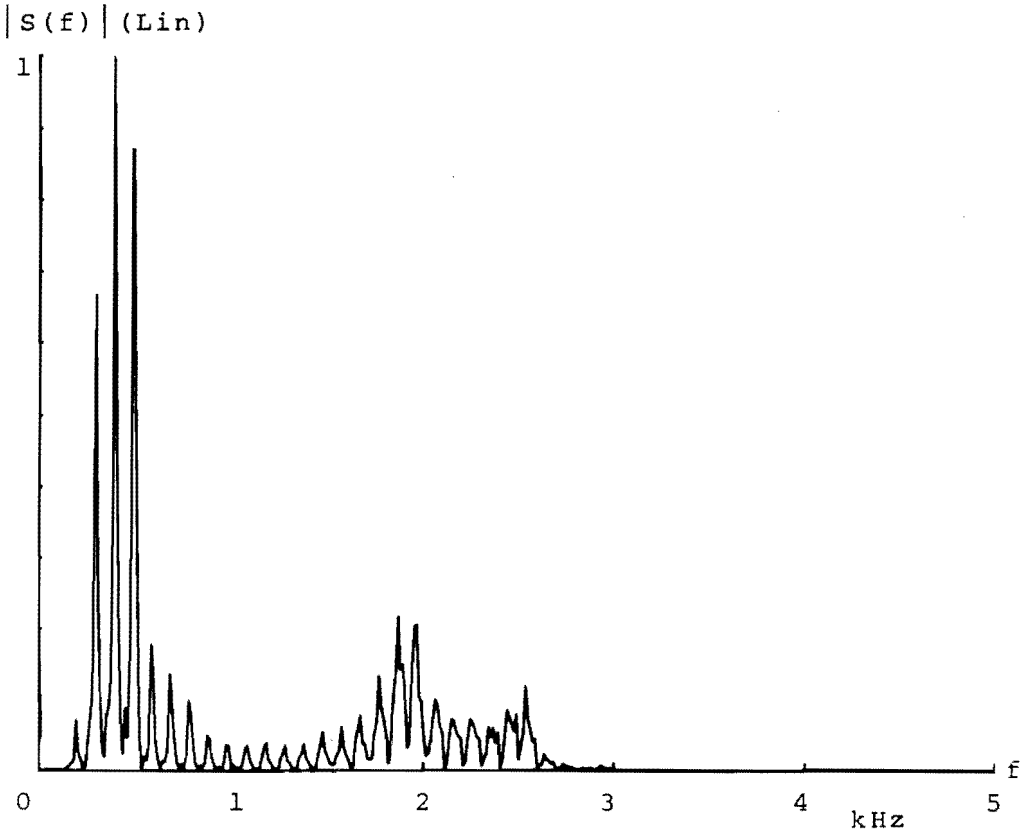


Fig 4.12.1

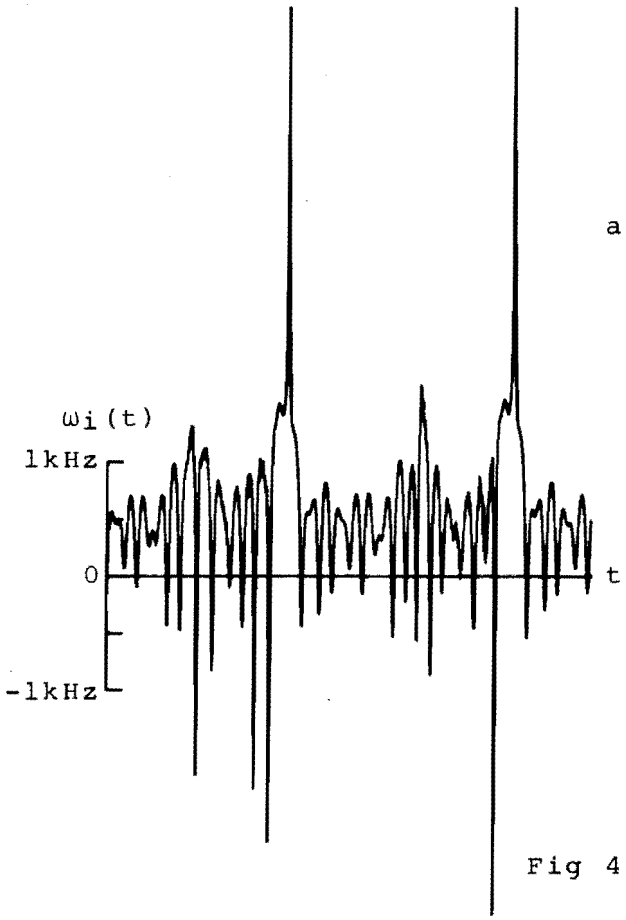


Fig 4.12.2

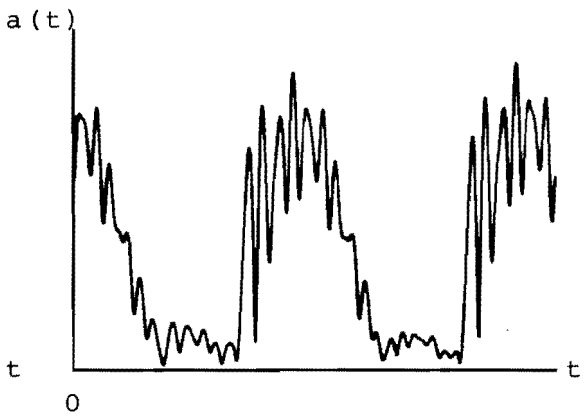


Fig 4.12.3

Figs 4.12.1-4.12.3 Analysis of Lowpass /ɛ/ (2400Hz)

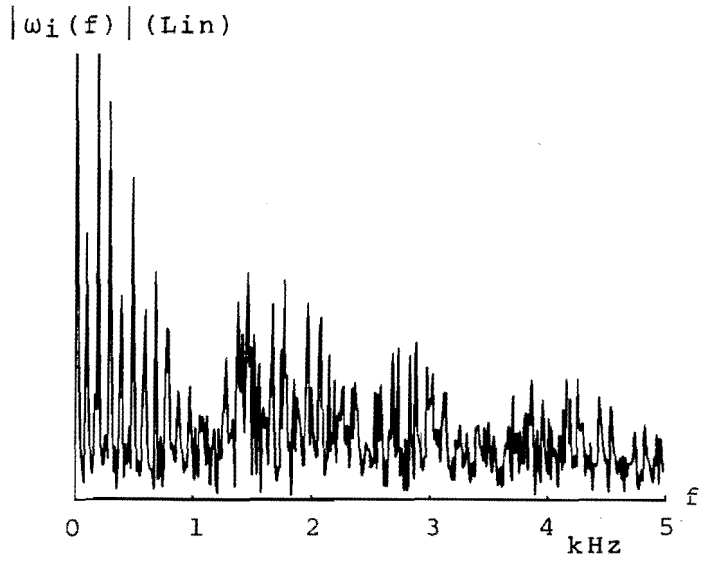


Fig 4.12.4

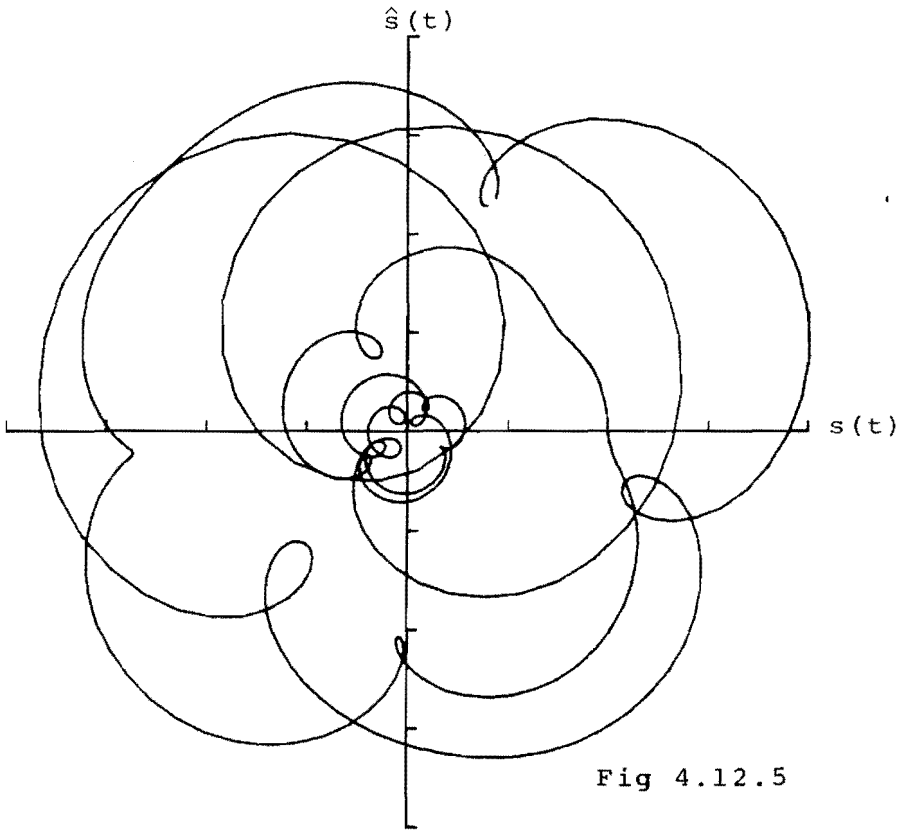


Fig 4.12.5

Figs 4.12.4-4.12.5 Analysis of Lowpass /ɛ/ (2400Hz)

The instantaneous amplitude waveform is affected less by the zero sign conversion and should display around 21 dips per cycle. There are only 17 countable dips per cycle superimposed on the 2 major UHP zero dips, but some minor ones may be inferred from the visible distortion of previously existing dips.

Reconsideration of the instantaneous waveforms of the full telephone bandwidth vowel, figure (4.4), reveals little change from this final low pass version. An increase in the magnitude of the third formant, however, has caused corresponding increases in the depth of some instantaneous waveform dips and the enlargement of some vector inner loops.

(4.3.1.2) FEMALE VOWEL ANALYSIS

Figure (4.13.1) is the time averaged amplitude spectrum of the vowel /ɛ/, this time uttered by a female speaker and low pass filtered to 3,400Hz. Noticeable features of the spectrum are the existence of a small DC component (due to slight A/D offset), a strong fundamental component modifying the vowel formant structure and the presence of only two identifiable formants. The fundamental frequency (line spacing) is approximately 200Hz and there are 12 significant harmonics.

A two cycle (10 ms) section of vowel waveform was chosen for analysis and figures (4.13.2) and (4.13.3) are the corresponding instantaneous functions. It is immediately apparent that neither instantaneous waveform is truly periodic, with instantaneous frequency indicating the conversion of an analytic signal complex zero from LHP to UHP in the interval between cycle 1 and cycle 2. Examination of the real and quadrature vowel time waveforms over the same time interval, figures (4.13.4) and (4.13.5), reveals no major aperiodicity, but confirms the zero conversions to UHP by the zero crossing count of both $s(t)$ and $\hat{s}(t)$ in cycle 2.

To aid in the analysis of the aperiodic instantaneous waveforms, vector loci were generated for each cycle. The vector plot for cycle 1, figure (4.13.6), shows only two full loops of the origin, indicating that average instantaneous frequency corresponds to the second harmonic. Consulting the amplitude spectrum, figure (4.13.1) leads us to expect two major UHP complex zeros per cycle, corresponding to the bandwidth between the second harmonic and the component at DC, and two major LHP zeros per cycle corresponding to the large 3rd and 4th harmonics. In total, the instantaneous frequency waveform for cycle 1 should display a basic shape of 2 major rises with approximately 10 dips superimposed. The instantaneous amplitude waveform should exhibit two major dips with 10 minor superimposed dips.

Examination of the instantaneous waveforms over cycle 1 reveals 12 superimposed instantaneous frequency dips and approximately 12 instantaneous amplitude dips or points of inflection. The presence of two additional significant LHP zeros indicates greater bandwidth than that predicted by the time averaged amplitude spectrum, figure (4.13.1)

The vector plot of cycle 2, figure (4.13.7), displays 3 full loops of the origin per cycle, again confirming the conversion of one zero to UHP. As average instantaneous frequency has stepped up to that of the 3rd harmonic, the amplitude spectrum now predicts one spike and 9 dips per cycle superimposed on the two instantaneous frequency rises, and 10 superimposed instantaneous amplitude dips or points of inflection per cycle.

The instantaneous waveforms appear to confirm the predictions for cycle 2, implying that the amplitudes of upper harmonics have fallen during this cycle. As the spectral content of cycle 2 fits the time averaged amplitude

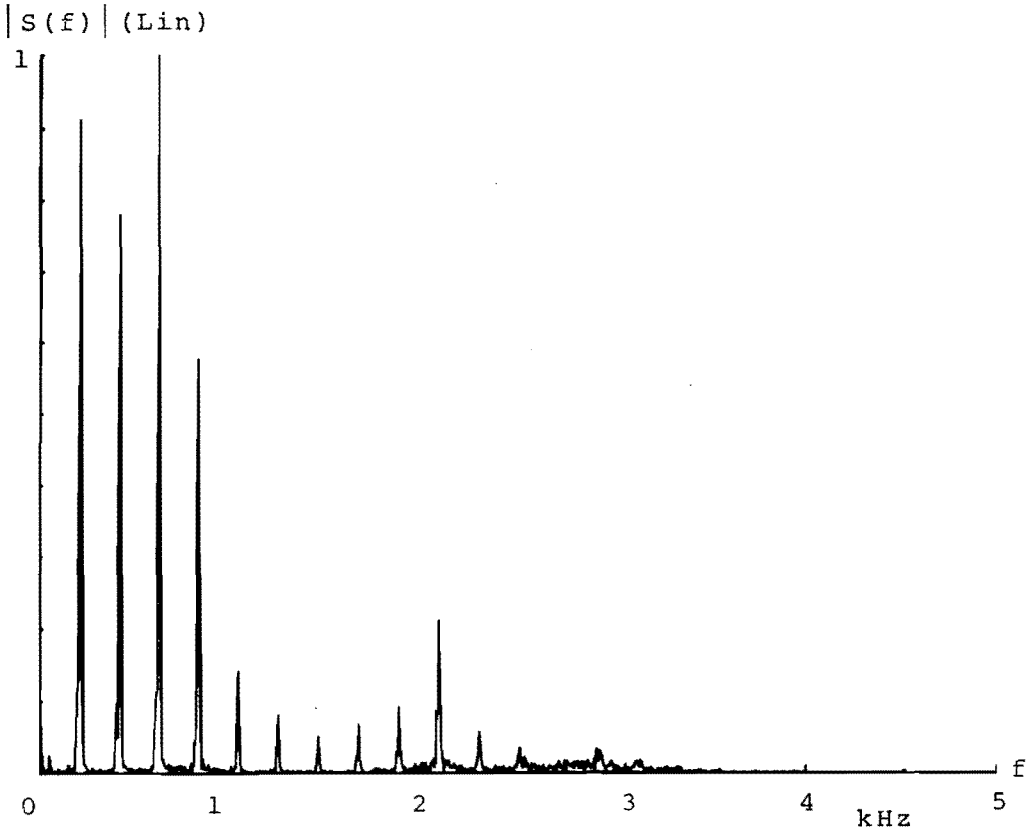


Fig 4.13.1

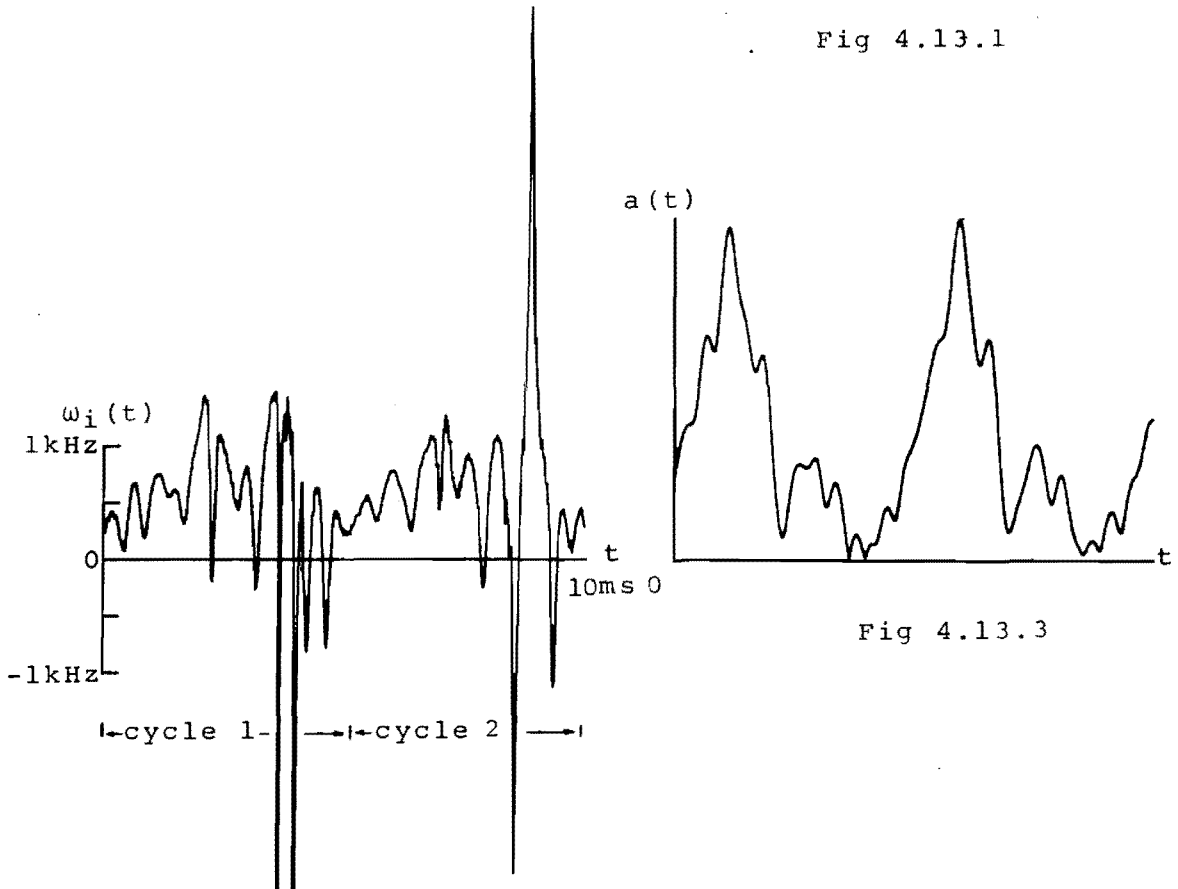


Fig 4.13.3

Fig 4.13.2

Figs 4.13.1 - 4.13.3 Analysis of Female /ɛ/

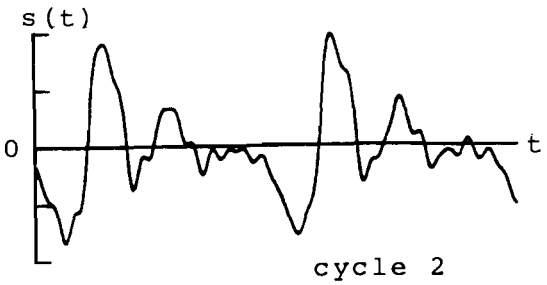


Fig 4.13.4

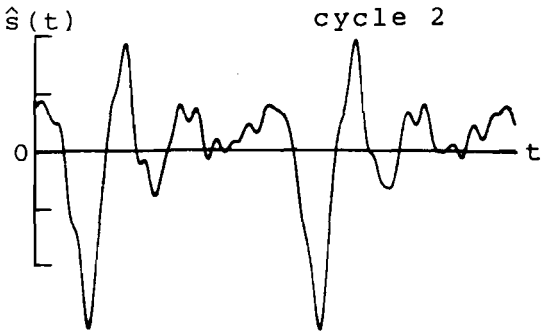


Fig 4.13.5

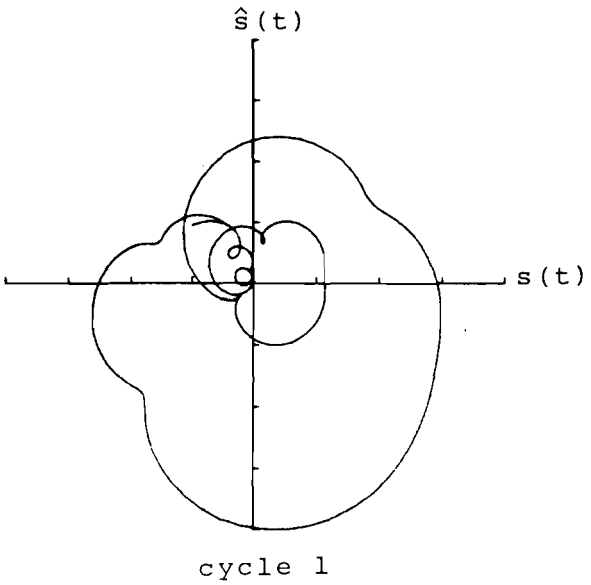


Fig 4.13.6

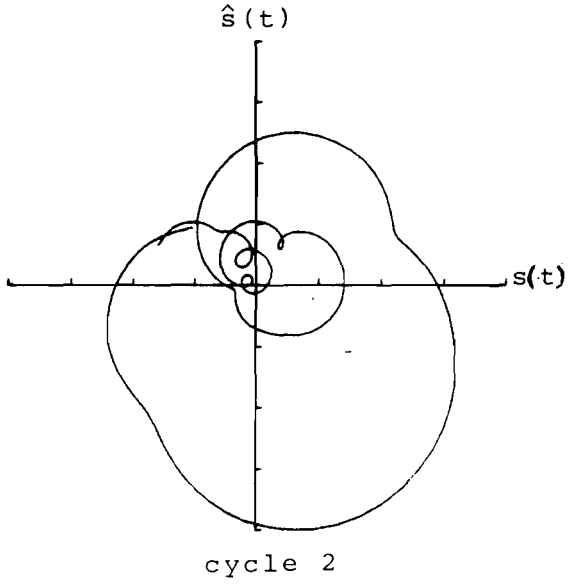


Fig 4.13.7

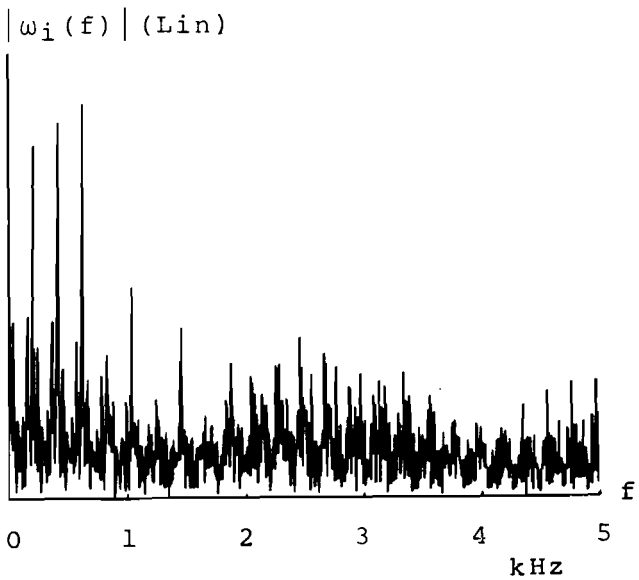


Fig 4.13.8

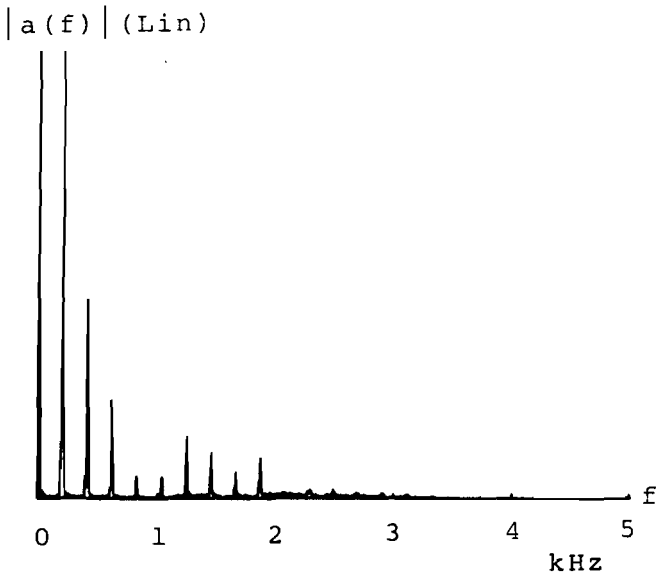


Fig 4.13.9

Figs 4.13.4-4.13.9 Analysis of Female /ɛ/

spectrum, it is possible to conclude that the vowel aperiodicity was due to changes in the magnitudes of some high frequency harmonic components.

Figures (4.13.8) and (4.13.9) are time averaged amplitude spectra of the instantaneous frequency and amplitude. Once again, instantaneous frequency displays a strong spectral line structure at low frequency, but is noise-like at high frequency and appears unbandlimited. Instantaneous amplitude, however, is periodic and band-limited and displays significant spectral energy at the difference frequency between the 1st and 2nd formants.

(4.3.1.3) SUMMARY

The following points have arisen from analysis of the instantaneous parameters of vowels.

(1) The average instantaneous frequency over one period of a vowel $\overline{\omega_i(t)}$, does not necessarily correspond to the frequency of the dominant spectral line, but is always a multiple of the line spacing frequency, ω (fundamental frequency).

(2) The "basic" cyclic shapes of vowel instantaneous waveforms are determined by the major spectral components (usually of the first formant). In both of the above examples, there have been two instantaneous frequency rises and two large instantaneous amplitude dips per cycle. These waveforms indicate two UHP analytic complex zeros per cycle and the "basic" instantaneous waveshapes for the male vowel are those generated by the first formant model, figures (4.7.1) and (4.7.2).

The "basic" waveforms exhibit an average instantaneous frequency, ω_1, which is not necessarily the same as $\overline{\omega_i(t)}$. The frequency ω_1 is related to the frequency of the lowest

spectral line in the bandpass spectrum by the relation

$$\omega_1 = (N_{UHP}+n_1).\omega \qquad \ldots (4.9)$$

where the lowest frequency spectral line is the n_1th harmonic of ω and N_{UHP} is the number of UHP analytic signal complex zeros indicated by the "basic" instantaneous frequency waveshape. If the spectrum is lowpass, $n_1=0$.

(3) In general, extension of the bandwidth of a signal not only increases the rate of occurrence of analytic signal zeros, but changes the positions of existing zeros. In the case of a vowel, however, it appears that as the upper harmonics are of low amplitude relative to the first formant, the effect of k additional analytic signal complex zeros per cycle on the instantaneous waveforms may be approximated by the superposition of k additional fluctuations per cycle.

The introduction of a second formant at frequency ω_{f2} causes the superposition of spike and/or dip fluctuations on the "basic" instantaneous waveforms at the rate $(\omega_{f2}-\omega_1)/\omega$ fluctuations per cycle. The frequency, ω_{f2}, is a multiple of ω and usually corresponds to the highest frequency major spectral component of the second formant.

The superimposed fluctuations are almost evenly spaced in time over one cycle of the fundamental frequency. The amplitudes of these fluctuations are generally greater during the periods of low average instantaneous amplitude (dictated by the "basic" instantaneous amplitude waveshape).

As the spectral magnitude of the second formant is often more than 10 dB below that of the first formant, the majority of superimposed instantaneous frequency fluctuations are dips corresponding to LHP analytic zeros. If, however, the second formant introduces N_s superimposed instantaneous frequency spikes per cycle, the average instantaneous frequency of the waveform becomes

$$\overline{\omega_i(t)} = \omega_1+\omega.N_s \qquad \ldots (4.10)$$

The second formant always causes $(\omega_{f2}-\omega_1)/\omega$ dips or dip-like fluctuations per cycle to be superimposed upon the "basic" instantaneous amplitude waveform. This explains the energy at frequency at $\omega_{f2}-\omega_1$ observed during spectral analysis of instantaneous amplitude waveforms.

(4) Introduction of the third formant, at frequency ω_{f3}, causes instantaneous waveform spikes and/or dips to occur at the rate $(\omega_{f3}-\omega_1)/\omega$ per cycle. Once again, these fluctuations are approximately evenly spaced in time, and most coincide with those previously generated by the second formant. As the spectral magnitude of the third formant is assumed to be less than that of the second formant, the introduced fluctuations are of low amplitude and principally cause distortion of the rise/dip pattern set up by the second formant.

If however, the third formant is of sufficient amplitude (but less than that of the second formant), $(\omega_{f3}-\omega_{f2})/\omega$, new instantaneous waveform fluctuations per cycle will become visible.

Equation (4.10) still applies for the vowel average instantaneous frequency, $\overline{\omega_i(t)}$, where N_s is now the total number of superimposed instantaneous frequency spikes introduced by formants two and three.

Spectral analysis of instantaneous amplitude now reveals energy at frequencies $\omega_{f2}-\omega_1$ and $\omega_{f3}-\omega_1$ and, due to the mechanism of superposition of dips, there may be a small component at $\omega_{f3}-\omega_{f2}$.

(4.3.2) UNVOICED FRICATIVE ANALYSIS

The characteristics of an unvoiced fricative which has been spectrally limited to the telephone bandwidth can be likened to those of bandlimited noise functions. For this reason it is convenient, and relevant, to preceed the analysis of real unvoiced fricatives with an examination of the probability density functions (pdf) and time waveforms of the instantaneous parameters of bandlimited Gaussian noise.

(4.3.2.1) GAUSSIAN NOISE

Expressions for the pdfs of the instantaneous parameters of rectangularly bandlimited Gaussian noise are presented in Section (2.3.3). Instantaneous amplitude is governed by the Rayleigh distribution and the pdf of instantaneous frequency, $p(\omega_t)$, is a symmetrical function around the noise band centre frequency ω_m. The spread of $p(\omega_t)$ is determined by the rectangular noise bandwidth $\Delta\omega$.

The spectral mean frequency of the noise signal, ω_m, is defined

$$\omega_m = \frac{\int_{-\infty}^{\infty} \omega . N(\omega) . d\omega}{\int_{-\infty}^{\infty} N(\omega) . d\omega} \quad \ldots (4.11)$$

where $N(\omega)$ is the signal spectral density. This can be shown to be equal to the average instantaneous frequency of the signal $\overline{\omega_i}$, where

$$\overline{\omega_i} = \lim_{T\to\infty} \frac{\int_{-T}^{T} \omega_i(t) . a^2(t) . dt}{\int_{-T}^{T} a^2(t) . dt} \quad \ldots (4.12)$$

(Ref. 112)

The result $\overline{\omega_i(t)} = \omega_m$ is predictable as $p(\omega_t)$ is symmetrical about ω_m.

The zero crossing rate of the noise waveform, λ_o, is expressed in terms of mean frequency and bandwidth by equation (4.13).

$$\lambda o = \frac{2}{\sqrt{12}\pi} \left\{ \frac{(\omega_m+\Delta\omega/2)^3 - (\omega_m-\Delta\omega/2)^3}{\Delta\omega} \right\}^{\frac{1}{2}} \quad \ldots (4.13)$$

The crossing rate, λo, is always greater than ω_m/π (the rate attributable to a sinusoid at frequency ω_m) when $\Delta\omega>0$. Additional zero crossings are caused by the analytic vector inner loops generated by negative instantaneous frequency excursions. Negative instantaneous frequencies are only possible when $\Delta\omega>0$ (Section (2.3.3)).

Gaussian noise, bandlimited to $\Delta\omega$, can be expected to exhibit $\Delta\omega/2\pi$ analytic signal complex zeros per second. The corresponding instantaneous waveforms will therefore display $\Delta\omega/2\pi$ fluctuations per second. The average instantaneous frequency, ω_m, and symmetrical nature of the noise bandwidth and $p(\omega_t)$ demands that half of the

analytic signal complex zeros be UHP corresponding to instantaneous frequency rises and half LHP, corresponding to dips.

From the work of Rice (Ref. 113), it is known that the expected rate of occurence of envelope maxima for a band-limited Gaussian process is

$$N_{MAX} = 0.6411 \ (\Delta\omega/2\pi) \qquad \ldots \ (4.14)$$

As this is also the rate of occurence of envelope minima (visible instantaneous amplitude dips), it serves as a prediction of the number of major analytic signal complex zeros per second. The rate is only 0.6411 times the average rate of analytic complex zero occurence.

(4.3.2.1.1) EXAMPLE (1)

The first example is the analysis of instantaneous functions generated by wide band Gaussian noise, limited to a 200 Hz bandwidth around a centre frequency of 3,000 Hz. Figure (4.14.1) is the logarithmic amplitude spectrum of a section of the noise waveform, 50 ms of which is plotted as figure (4.14.2). The corresponding instantaneous functions are plotted in figures (4.14.3) and (4.14.4).

From Rice's formula for the rate of envelope maxima, equation (4.14), it is possible to predict that major analytic signal complex zeros will occur for this signal at around 128 per second. This corresponds to approximately 6.4 major instantaneous parameter fluctuations over the displayed 50 ms segment.

There are approximately 4 visible instantaneous amplitude minima over the time interval and these correspond to instantaneous frequency fluctuations.

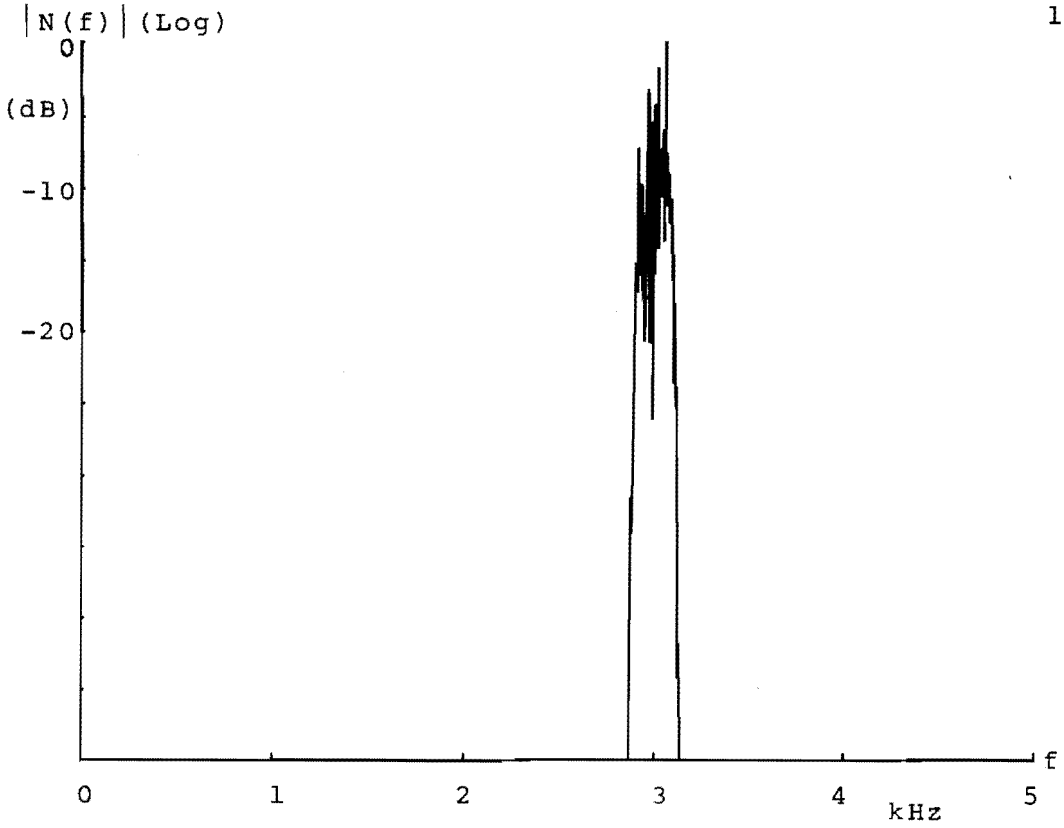


Fig 4.14.1

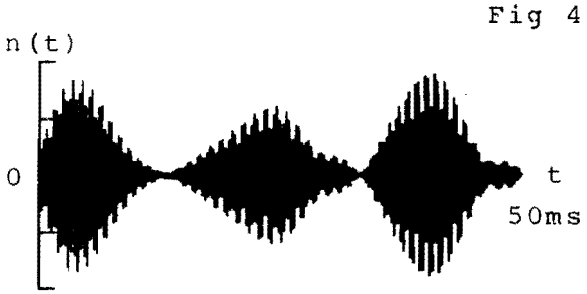


Fig 4.14.2

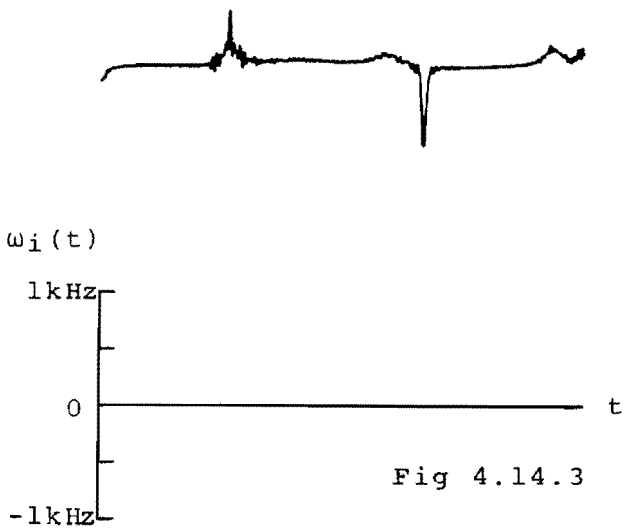


Fig 4.14.3

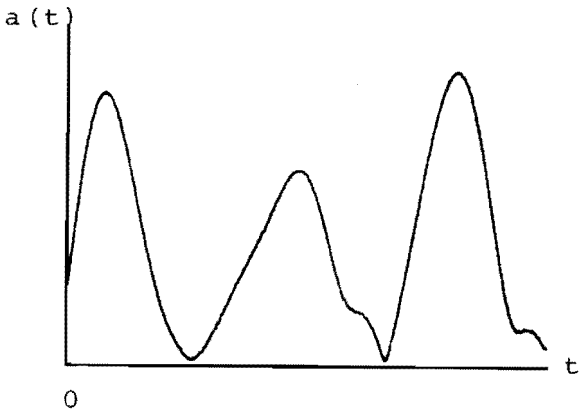


Fig 4.14.4

Figs 4.14.1-4.14.4 Narrow Band Noise Analysis (200Hz)

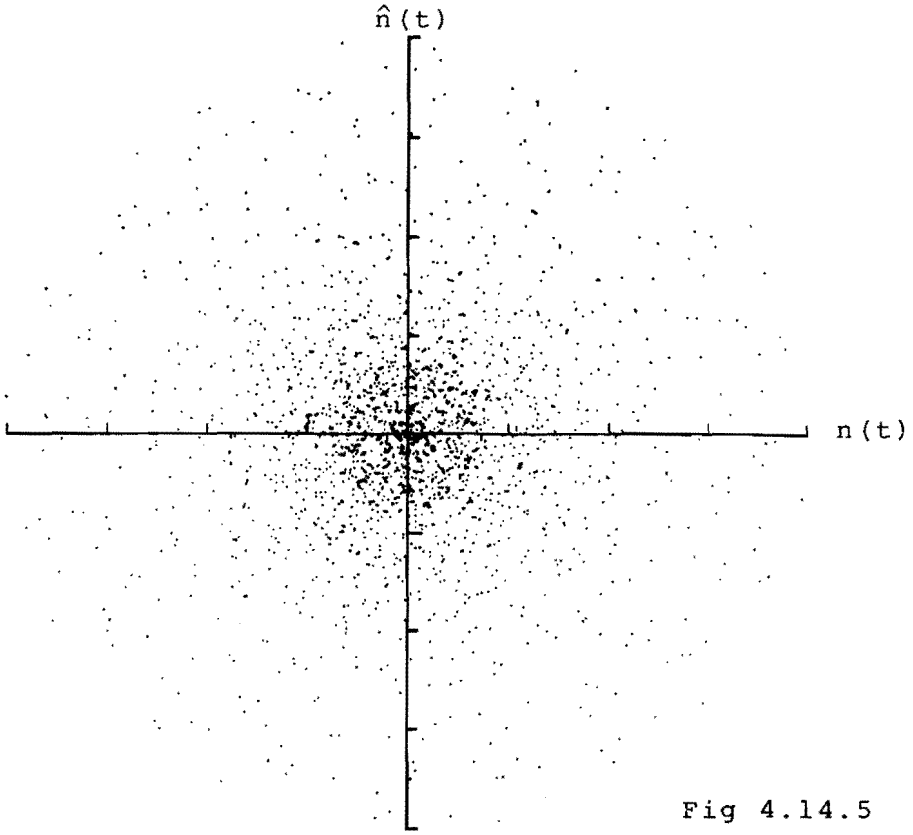


Fig 4.14.5

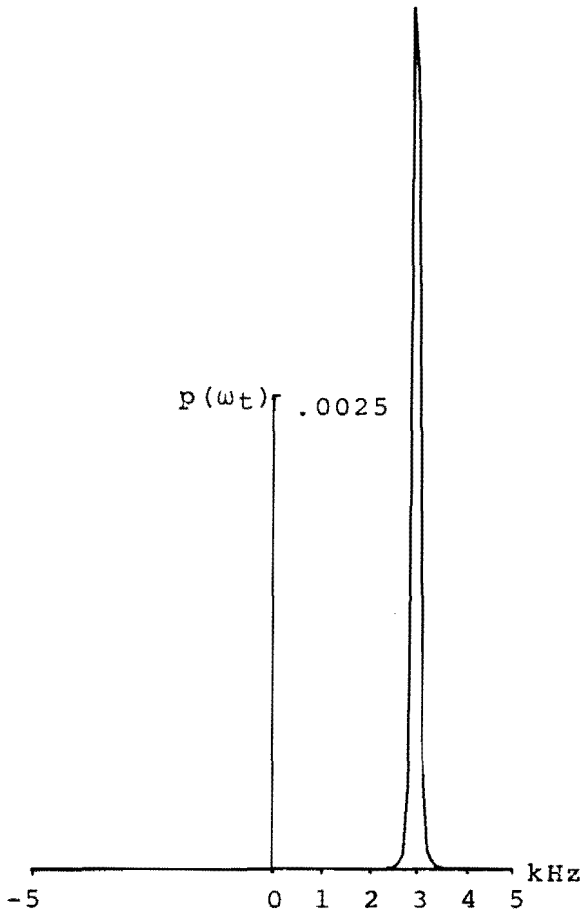


Fig 4.14.6

Fig 4.14.5-4.14.6 Narrow Band Noise Analysis (200Hz)

One UHP complex zero and one LHP complex zero are clearly indicated in figure (4.14.3), with smaller fluctuations also visible.

The plot of 2,000 vector samples, figure (4.14.5), gives an indication of the instantaneous amplitude distribution. The pdf is known to be Rayleigh and its peak appears to be close to the origin.

The pdf of instantaneous frequency, figure (4.14.6), has been generated by dividing the frequency axis (-5,000 Hz to +5,000 Hz) into 100 Hz segments and registering the number of instantaneous frequency samples falling in each. In order to obtain the smoothest approximation to the pdf, all 14,880 available instantaneous frequency samples have been used.

This plot confirms that $\overline{\omega_i(t)} \equiv 3{,}000$ Hz and shows $p(\omega_t)$ to be symmetrical about 3,000 Hz (within the limits of frequency resolution).

(4.3.2.1.2) EXAMPLE (2)

The second example is wideband Gaussian noise, this time bandlimited to 1,000 Hz around a 3,000 Hz centre frequency. Figure (4.15.1) is the corresponding logarithmic amplitude spectrum. The time waveform is illustrated by a 50 ms segment, figure (4.15.2), and the derived instantaneous waveforms are figures (4.15.3) and (4.15.4).

Rice's equation now predicts 32.1 instantaneous amplitude dips over the 50 ms period, and there are 32 minima visible in figure (4.15.4). As usual, the major instantaneous amplitude dips correspond to major instantaneous frequency fluctuations. Spike and dip fluctuations appear to be evenly distributed around $\overline{\omega_i(t)}$.

The presence of several negative instantaneous frequency excursions over the 50 ms period, indicates

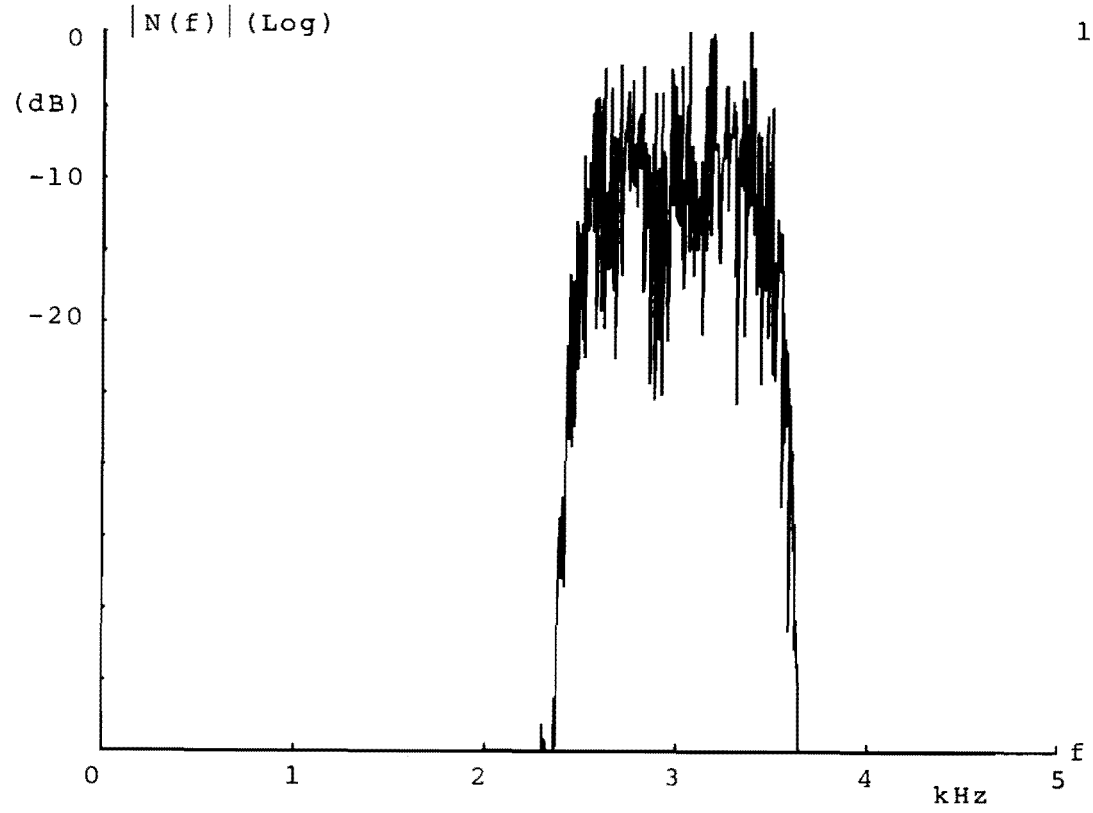


Fig 4.15.1

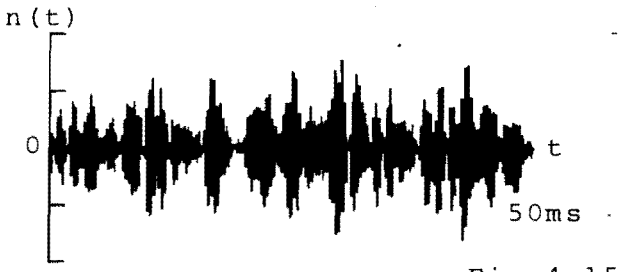


Fig 4.15.2

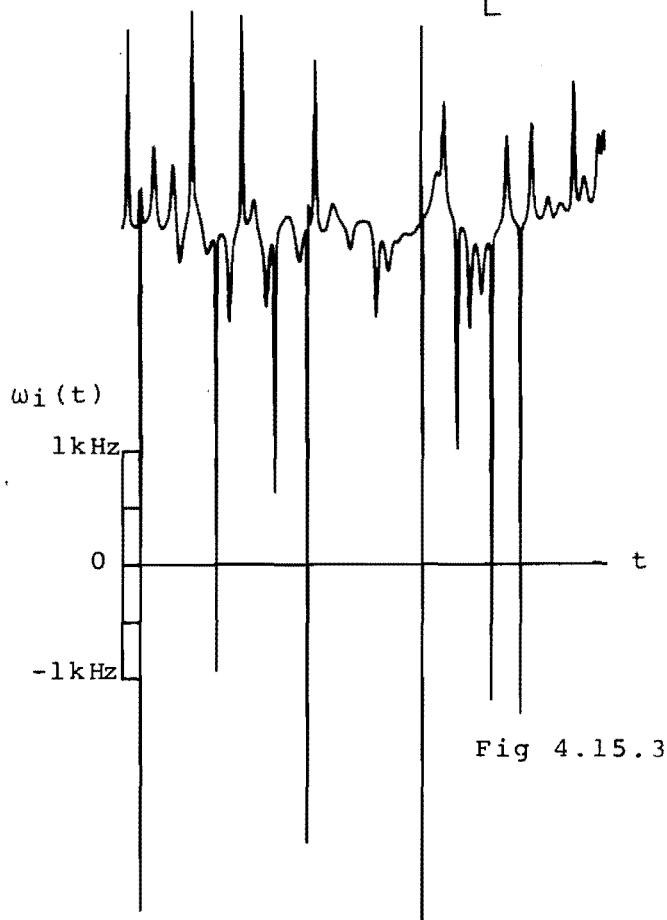


Fig 4.15.3

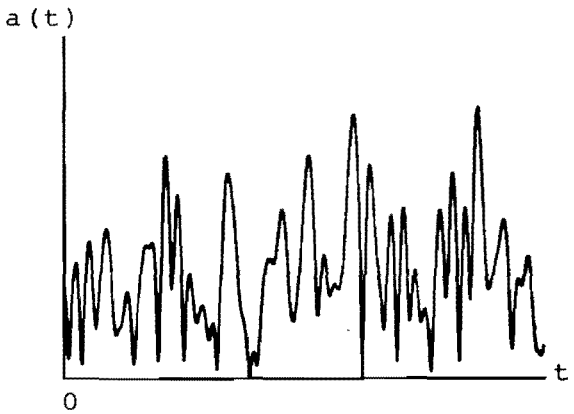


Fig 4.15.4

Figs 4.15.1-4.15.4 Narrow Band Noise Analysis (1000Hz)

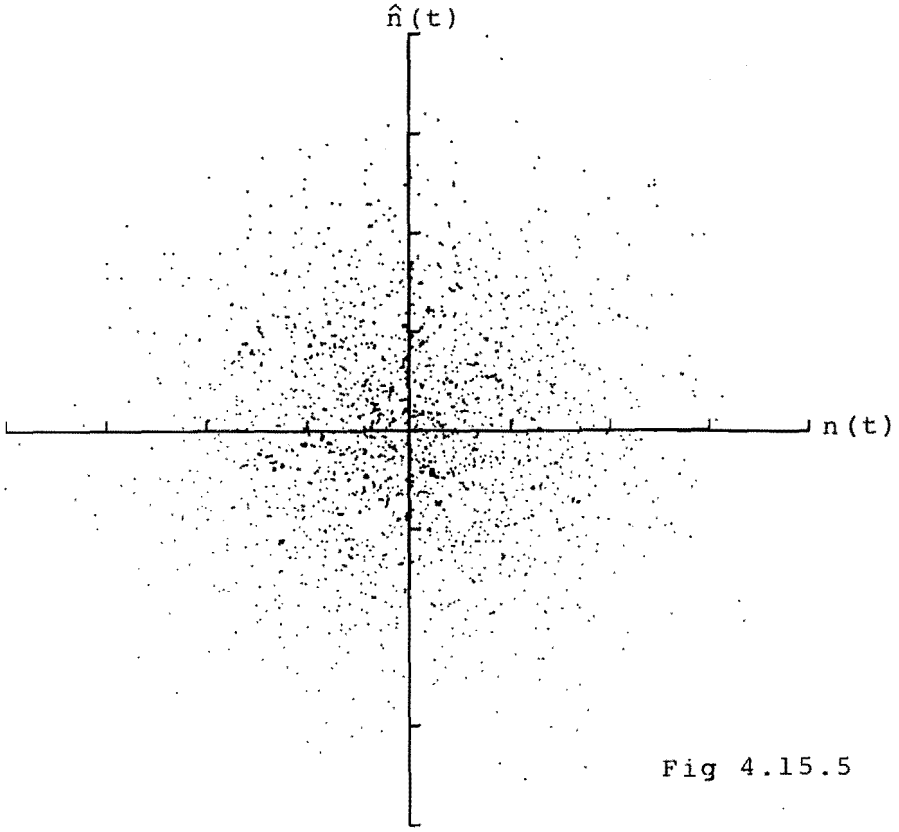


Fig 4.15.5

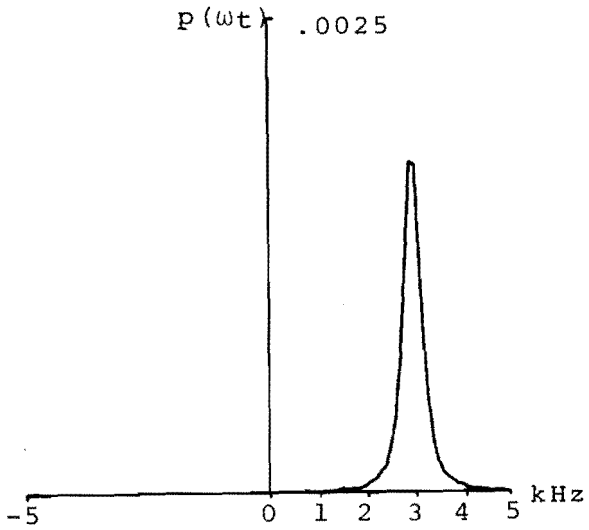


Fig 4.15.6

Figs 4.15.5-4.15.6 Narrow Band Noise Analysis (1000Hz)

the existence of analytic vector inner loops and suggests that the zero crossing rate may be significantly higher than $\omega_m/2\pi$ = 6,000 per second. The formula for crossing rate, equation (4.13), yields λo = 6028 per second, meaning that a zero crossing based estimate of the centre frequency would be 14 Hz too high.

Figure (4.15.5) is the plot of 2,000 vector samples which, once again, reflects the Rayleigh amplitude distribution.

The pdf of instantaneous frequency, figure (4.15.6) is plotted from 15,280 available samples. It is symmetrical around 3,000 Hz, but shows a wider spread than figure (4.14.6).

The usefulness of the instantaneous frequency pdf as a tool for spectral analysis can be illustrated using figure (4.15.6). The probability of an instantaneous frequency sample falling in the frequency range which corresponds to the pdf maximum is given by

$$p(\omega_i(kT) = \overline{\omega_i(t)}) = (N_m/200\pi)/N \quad . \; . \; (4.15)$$

where N_m is the number of samples in pdf maximum frequency range and N is the total number of samples. In the above case, N_m=2,652 and N=15,280, yielding $p(\omega_i(kT)=\overline{\omega_i(t)})=2.76\times10^{-6}$.

As the noise bandwidth is approximately rectangular, then

$$p(\omega_i(kT)=\overline{\omega_i(t)}) = {}^1/_2(\sqrt{12}/\Delta\omega) \quad . \; . \; . \; (4.16)$$

from Section (2.3.3). Using the value of $p(\omega_i(kT)=\overline{\omega_i(t)})$ obtained from equation (4.15) and rearranging equation (4.16) gives $\Delta\omega=2\pi\times999$ radians per second. This result is in excellent agreement with prefiltering of the wideband noise source for this example.

(4.3.2.2) UNVOICED FRICATIVE /s/

Having developed techniques for analysing the instantaneous parameters of bandlimited Gaussian noise, it is now possible to apply these to an unvoiced fricative.

Figure (4.16.1) is the time averaged logarithmic amplitude spectrum of the unvoiced fricative /s/ uttered by a male speaker and low pass filtered to 3,400 Hz. The plot, figure (4.16.2), is a 10 ms segment of the time waveform and figures (4.16.3) and (4.16.4) are the corresponding instantaneous parameter plots. The amplitude spectrum suggests that the low pass fricative is principally bandpass noise centred at 3,300 Hz with a smaller noise spectral peak at 2,400 Hz.

The vector plot, figure (4.16.5) is plotted from the 10 ms of displayed instantaneous data. It is equivalent to the dot plots, figures (4.14.5) and (4.15.5), except that in this case, consecutive dots have been joined by straight lines. This process reveals the presence of vector inner loops and loops which only just encircle the origin. The vector amplitude distribution appears similar to the Rayleigh.

The pdf of instantaneous frequency, figure (4.16.6), is plotted from 8,939 available samples and is almost symmetrical around $\omega_m/2\pi$=3,250Hz. If it is assumed that this pdf was obtained from rectangularly filtered Gaussian noise, an estimate of the filter bandwidth can be calculated from the distribution maximum. Substituting the values N_m=758 and N=8,939 into equation (4.15) and using equation (4.16) yields the equivalent rectangular bandwidth $\Delta\omega=2\pi$x2,042 radians per second.

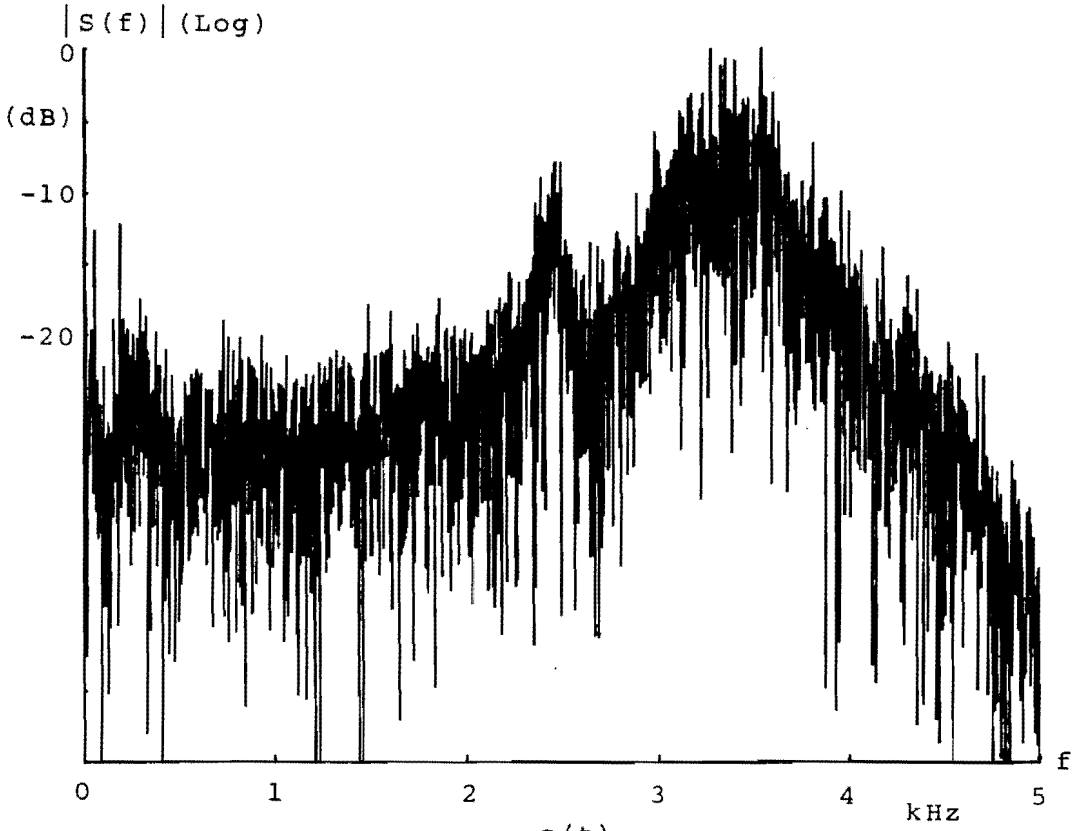


Fig 4.16.1

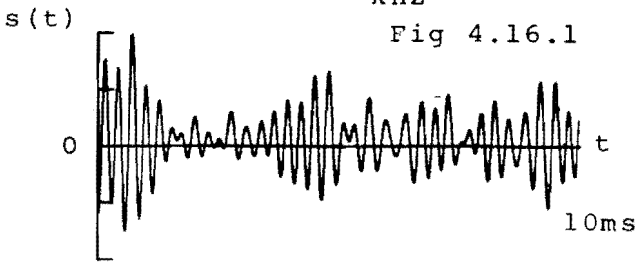


Fig 4.16.2

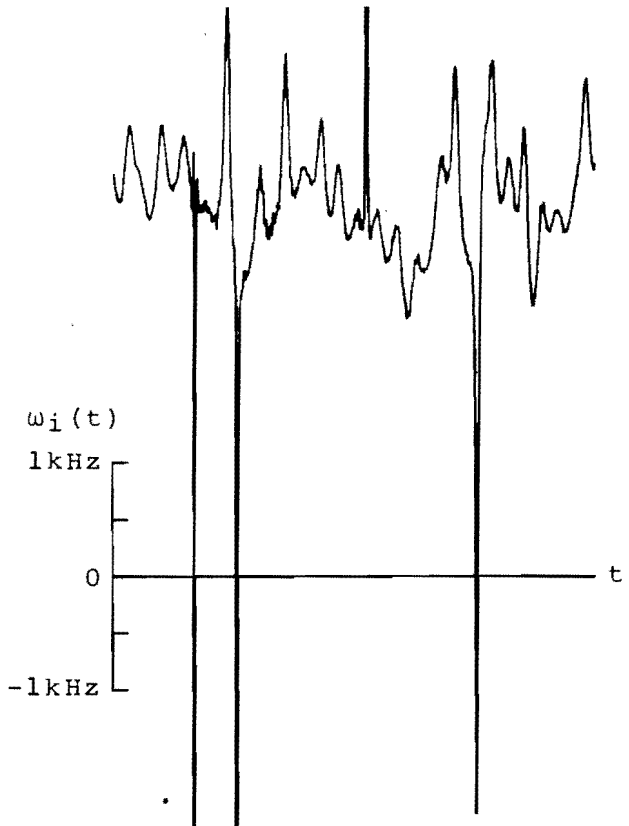


Fig 4.16.3

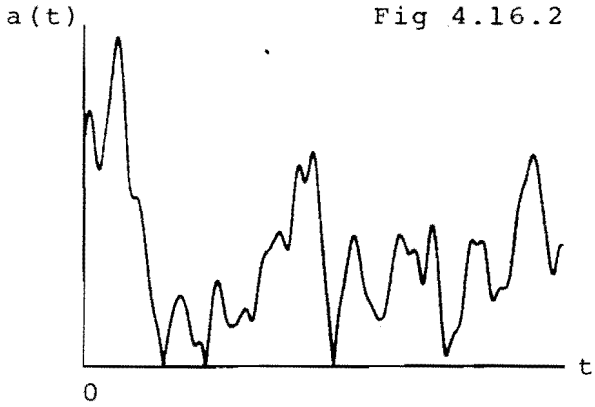


Fig 4.16.4

Figs 4.16.1-4.16.4 Analysis of /s/

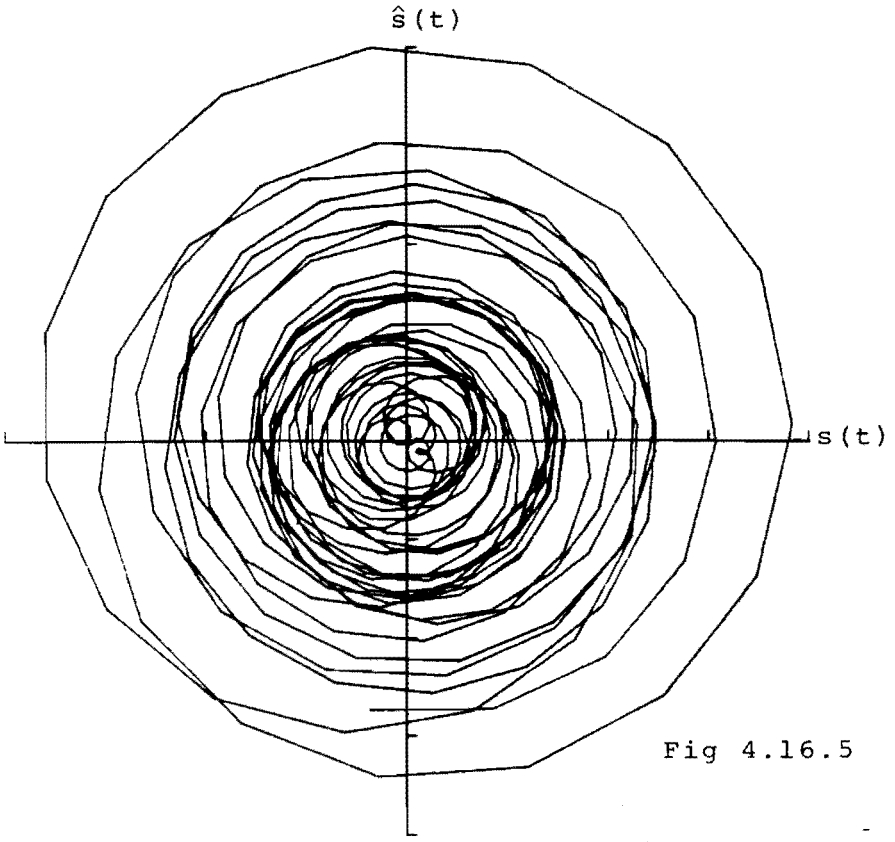


Fig 4.16.5

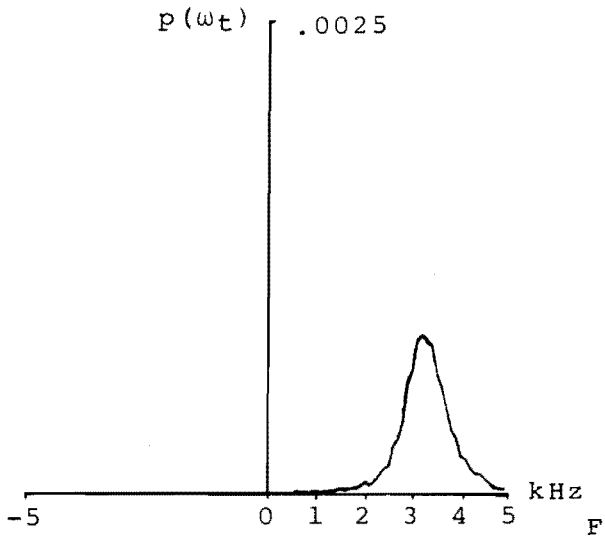


Fig 4.16.6

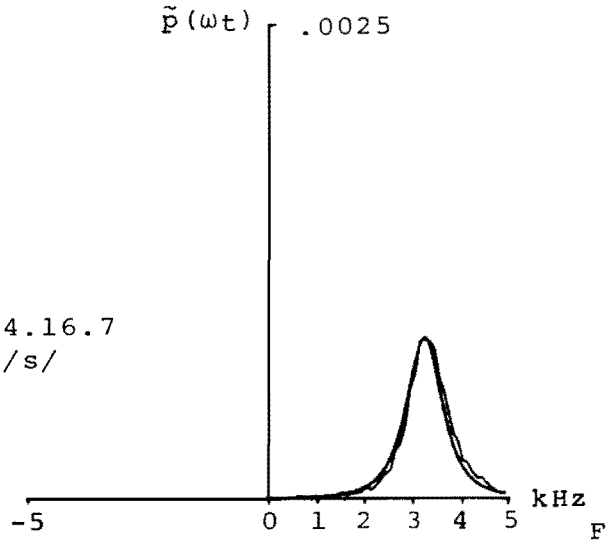


Fig 4.16.7

Figs 4.16.5-4.16.7
Analysis of /s/

Plotting the pdf of rectangularly filtered Gaussian noise $\omega_m=2\pi x3,250$, $\Delta\omega=2$ x2,000 on top of the pdf obtained from the unvoiced phoneme produces the composite plot, figure (4.16.7). This shows that the spread of instantaneous frequency values is very similar in both cases.

Substituting the equivalent rectangular bandwidth, $\Delta\omega$, into Rice's equation for the rate of envelope maxima (or minima) suggests that instantaneous amplitude should exhibit approximately 13 minima over the 10 ms period. Examination of the instantaneous amplitude waveform, figure (4.16.3), reveals 16 or 17 dips in 10 ms. Including dip distortions, there are 23 discernable instantaneous amplitude fluctuations in figure (4.16.3) and this points to the presence of more than 20 analytic signal complex zeros in a 10 ms period. For this to occur, the effective signal bandwidth must have been greater than 2,000 Hz over the 10 ms time interval.

(4.3.2.3) UNVOICED FRICATIVE /ʃ/

In order to confirm the usefulness of Gaussian noise analysis techniques when applied to stationary unvoiced fricatives, a similar analysis to the above has been performed on /ʃ/, uttered out of context by a male speaker. Figure (4.17.1) is the fricatives logarithmic amplitude spectrum. A 30 ms section of the time waveform is plotted in figure (4.17.2) and figures (4.17.3) and (4.17.4) are the corresponding instantaneous waveforms.

The pdf of instantaneous frequency figure (4.17.5) has been plotted from 15,369 available samples and appears to be symmetrical about a centre frequency of 2,350 Hz. Substituting the values N_m=1,612 and N=15,369 into equation (4.15) and using equation (4.16) yields the equivalent Gaussian rectangular bandwidth $\Delta\omega=2\pi x1,650$ radians per second.

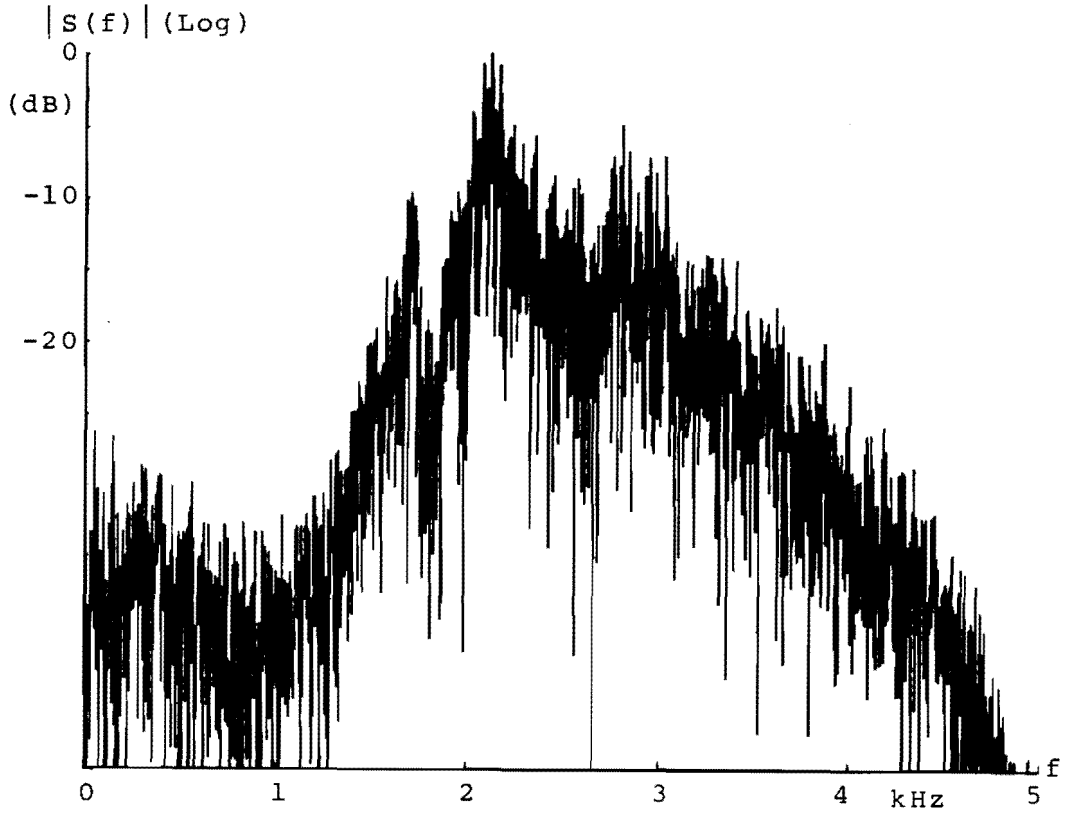


Fig 4.17.1

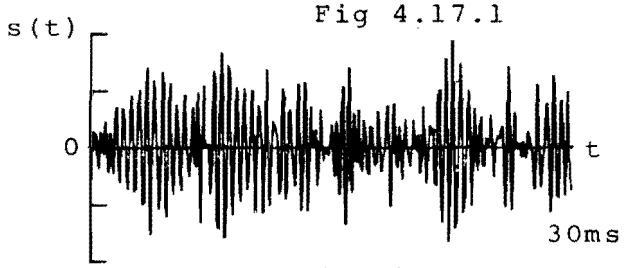


Fig 4.17.2

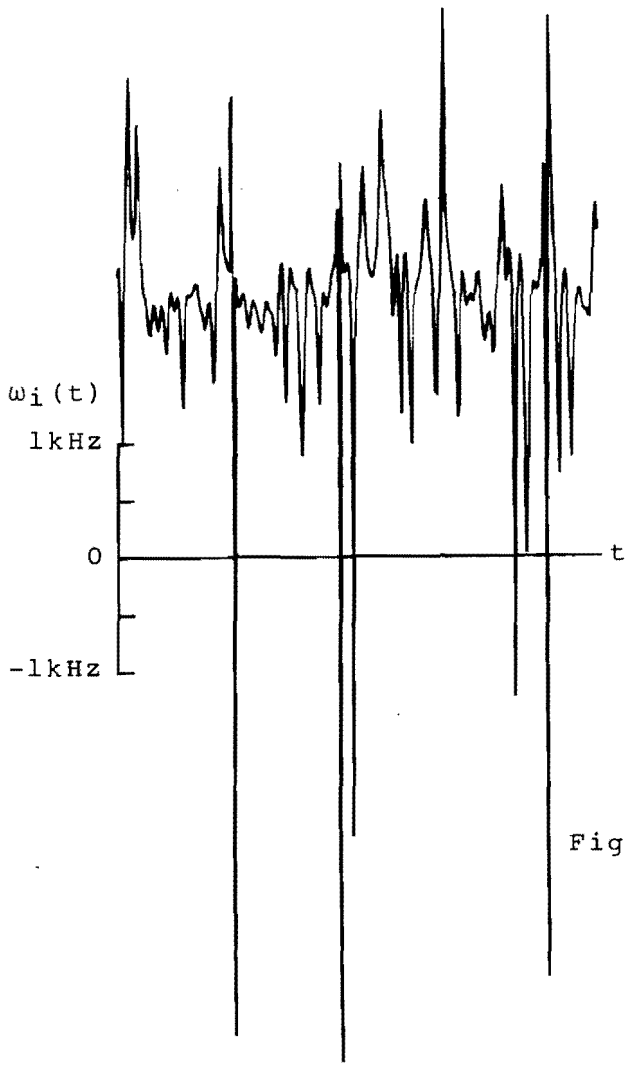


Fig 4.17.3

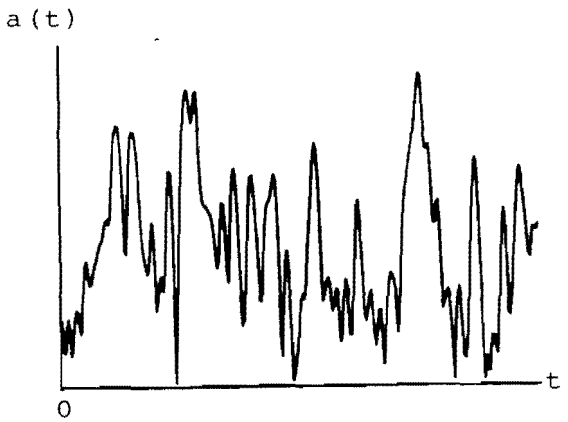


Fig 4.17.4

Figs 4.17.1-4.17.4 Analysis of /ʃ/

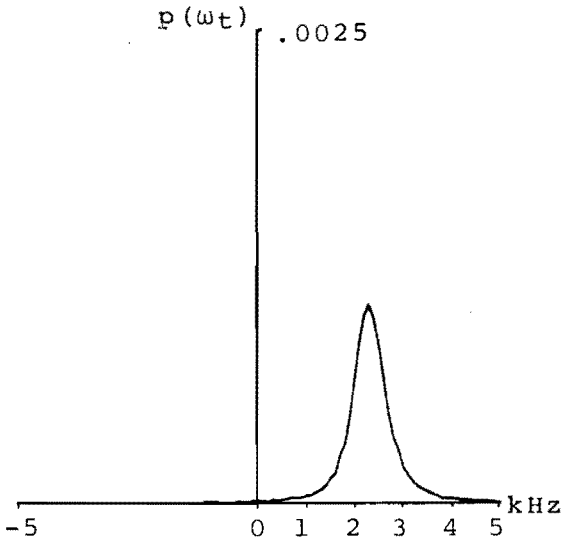


Fig 4.17.5

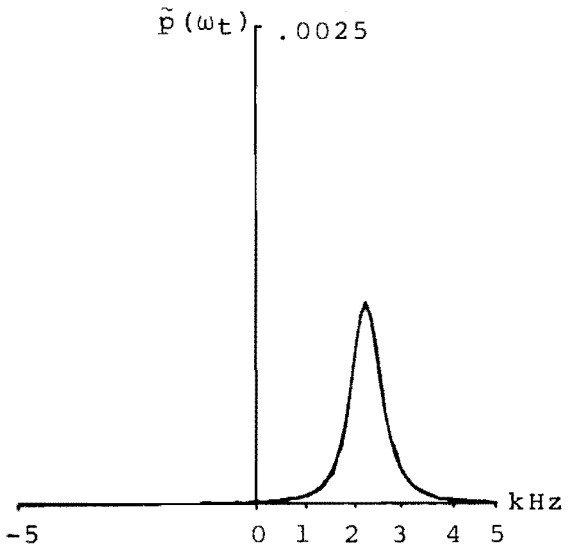


Fig 4.17.6

Figs 4.17.5-4.17.6 Analysis of /ʃ/

Calculating the equivalent Gaussian signal pdf for $\omega_m=2\pi x2,350$ and the above value of $\Delta\omega$, and superimposing this upon the pdf of /ʃ/ yields the composite plot, figure (4.17.6). It can be seen that the distributions match well.

Rice's formula for the rate of envelope maxima, equation (4.14), predicts 32 instantaneous amplitude dips over a 30 ms period for a Gaussian signal with the above value of $\Delta\omega$. The count of 37 visible minima in figure (4.17.4) does not differ greatly from the prediction, but indicates that the fricative is probably not fully described by a bandpass Gaussian representation.

(4.3.2.4) UNVOICED FRICATIVE RECONSTRUCTIONS

The amplitude spectra of unvoiced fricatives are known to be the result of vocal tract resonances and antiresonances. This means that an equivalent rectangular bandwidth and centre frequency is not a precise model, but it may be a suitable representation in the case of fricatives limited to the telephone bandwidth.

In order to test the validity of this representation for the fricatives /s/ and /ʃ/, reconstructions were made using appropriately bandlimited Gaussian noise ($\omega_m=2\pi x3,250$, $\Delta\omega=2\pi x2,000$ for /s/ and $\omega_m=2\pi x2,350$, $\Delta\omega=2\pi x1,640$ for /ʃ/). Listening tests proved that the noises were similar to, but easily distinguished from the original fricatives.

Further tests showed that the reconstructions could be made to sound more realistic by adding wideband Gaussian noise (at approximately -20 dB) and lowpass filtering the resulting signals to 3,400 Hz. Although still distinguishable from the original fricatives, these improved reconstructions could be recognised as /s/ and /ʃ/ when listened to out of context.

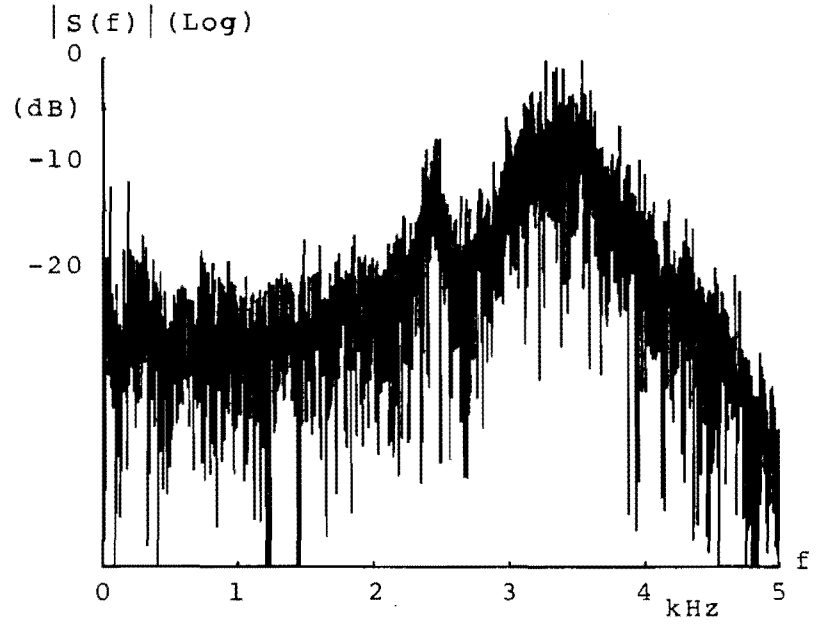


Fig 4.18.1

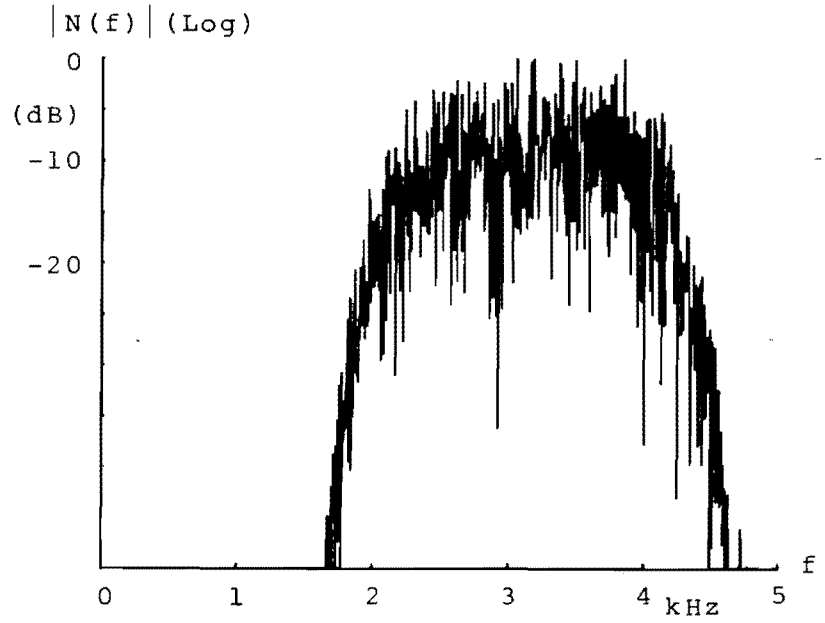


Fig 4.18.2

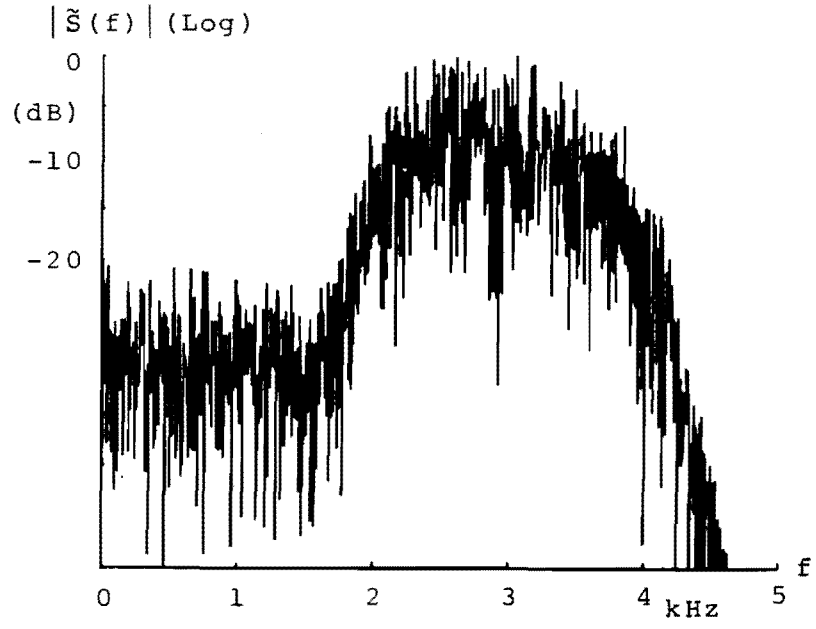


Fig 4.18.3

Figs 4.18.1-4.18.3 Reconstruction of /s/

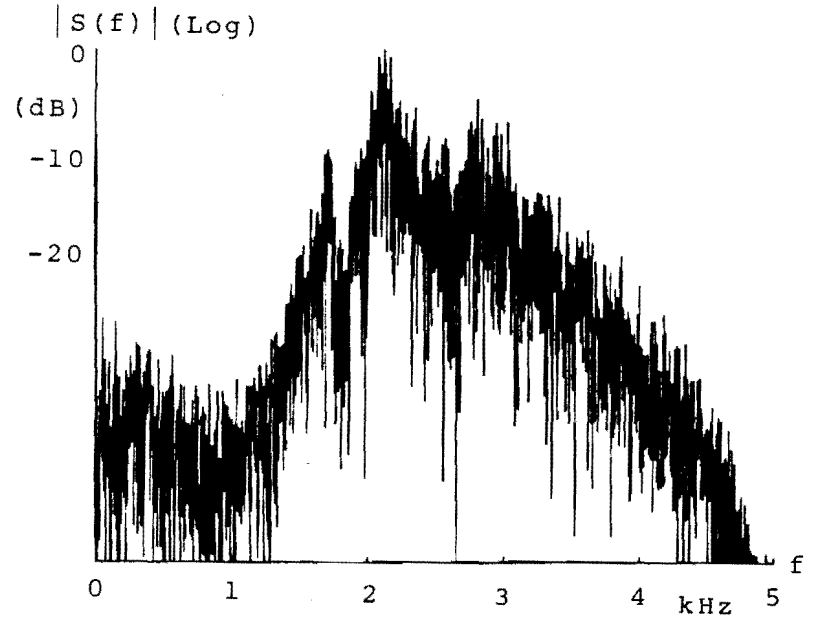


Fig 4.18.4

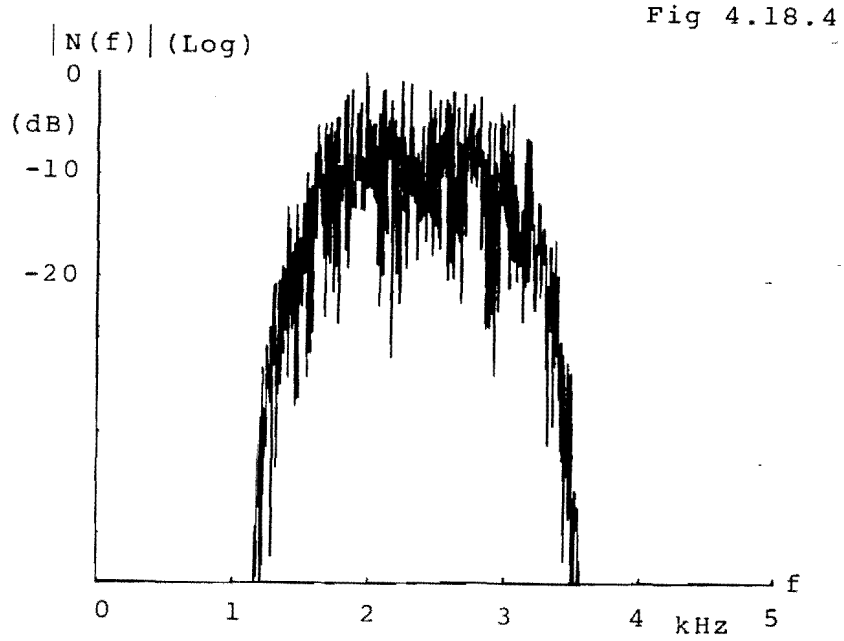


Fig 4.18.5

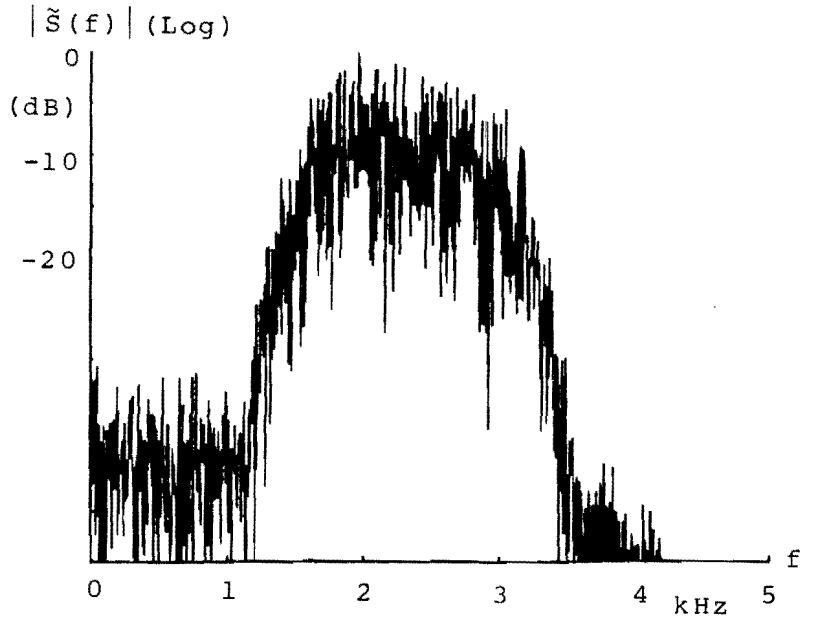


Fig 4.18.6

Figs 4.18.4-4.18.6 Reconstruction of /ʃ/

Figure (4.18) illustrates the reconstruction process for both fricatives. Figure (4.18.1) is the amplitude spectrum of /s/, figure (4.18.2) the basic rectangular bandwidth reconstruction and figure (4.18.3) the improved reconstruction. Figure (4.18.4) is the amplitude spectrum of /ʃ/ and figures(4.18.5) and (4.18.6) the basic and improved reconstructions.

The real and artificial sounds used to generate figures (4.18.1) to (4.18.6) are presented as Part 1, of the cassette tape which accompanies this thesis.

Success with listening tests suggests that an equivalent rectangular bandwidth and centre frequency is a suitable model for bandlimited unvoiced fricatives.

(4.3.3) VOICED FRICATIVE ANALYSIS

Voiced fricatives are generated by the same vocal tract constrictions which produce unvoiced fricatives, but in this case air flow from the glottis is voiced. The resultant amplitude spectra display strong spectral lines at low frequency and low amplitude noise at high frequency.

Before examining the instantaneous parameters of a voiced fricative, it is convenient to consider a model constructed from a single low frequency sinusoid plus noise.

(4.3.3.1) EXAMPLE

The signal, s(t), is a voiced fricative model which consists of a constant amplitude sinusoid at frequency $\omega_s=2\pi x500$ radians per second and Gaussian noise, n(t), with approximate rectangular bandwidth $\Delta\omega=2\pi x1,000$ radians per second and mean frequency $\omega_m=2\pi x3,000$ radians per second.

$$s(t)=A\cos\omega_s t+n(t) \qquad \ldots (4.17)$$

The ratio of sinusoid and noise amplitudes A/σ, where σ is the square root of the mean noise power, is set to $A/\sigma=2.5$. Figure (4.19.1) is the amplitude spectrum of s(t) and 50 ms of the time waveform is illustrated in figure (4.19.2). The corresponding instantaneous waveforms are figures (4.19.3) and (4.19.4).

The average instantaneous frequency is 500 Hz, corresponding to the frequency of the sinusoid. As this is the lowest frequency in the bandpass spectrum, all analytic signal zeros will be LHP meaning that all instantaneous frequency fluctuations must be dips. The signal bandwidth is 3,000 Hz, but the difference frequency between the sinusoid and noise mean frequency is $(\omega_m-\omega_s)/2\pi=2,500$ Hz. LHP analytic signal zeros will therefore occur at the rate of 3,000 per second, but "significant" instantaneous frequency and amplitude dips should be generated at 2,500 per second. This was confirmed using rough counting techniques on figures (4.19.3) and (4.19.4).

The vector amplitude distribution, figure (4.19.5), now conforms to a Rician Distribution, equation (4.18), in this case for $A/\sigma=2.5$.

$$p(a_t)=\frac{a_t}{\sigma^2}\exp\left(-\frac{a_t^2+A^2}{2\sigma^2}\right)I_o\left(\frac{A.a_t}{\sigma^2}\right) \qquad \ldots (4.18)$$

I_o is the modified Bessel function of the first kind and zero order. A cross section of the Rician distribution is illustrated in figure (4.20) for A/σ in the range 0 to 5. When $A/\sigma=0$, the distribution reverts to the Rayleigh. When the value of A is large with respect to σ, however,

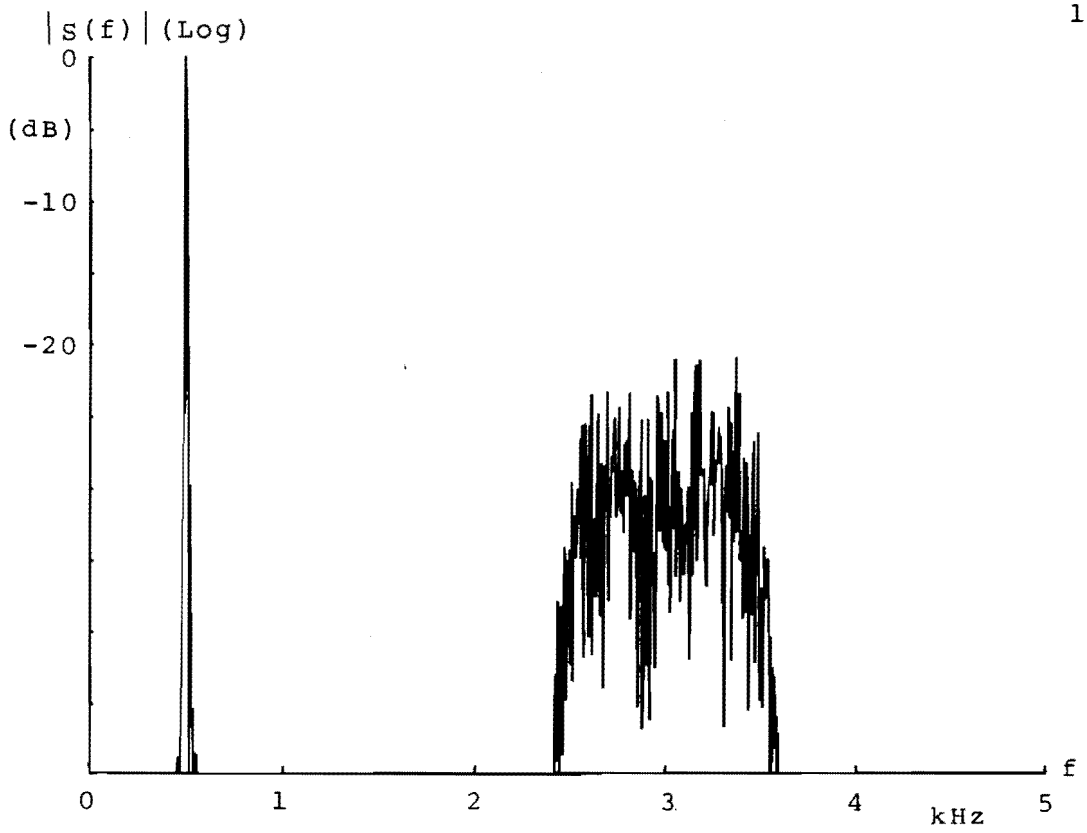


Fig 4.19.1

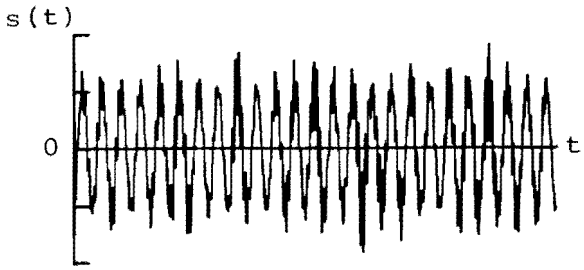


Fig 4.19.2

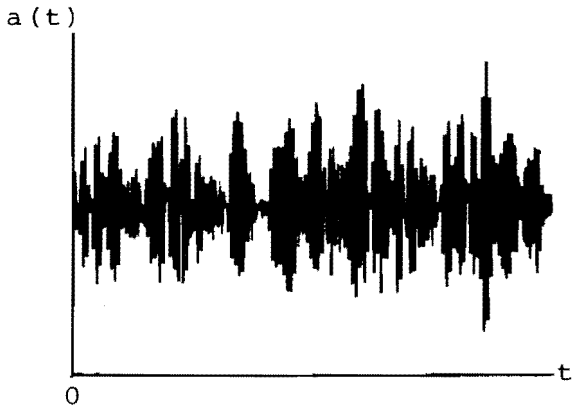


Fig 4.19.4

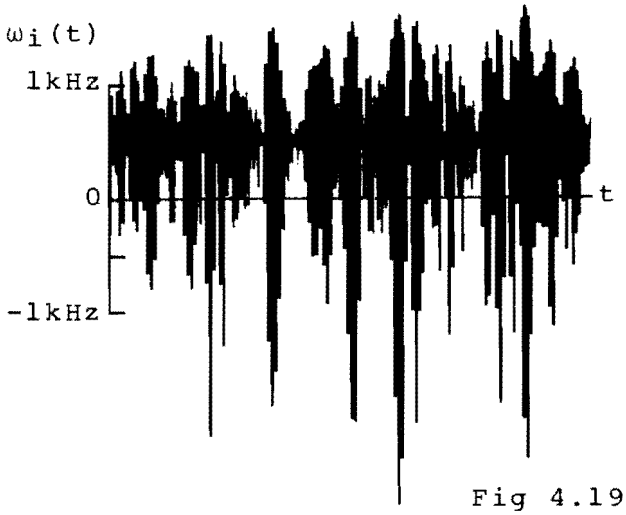


Fig 4.19.3

Figs 4.19.1-4.19.4 Voiced Fricative Model Analysis

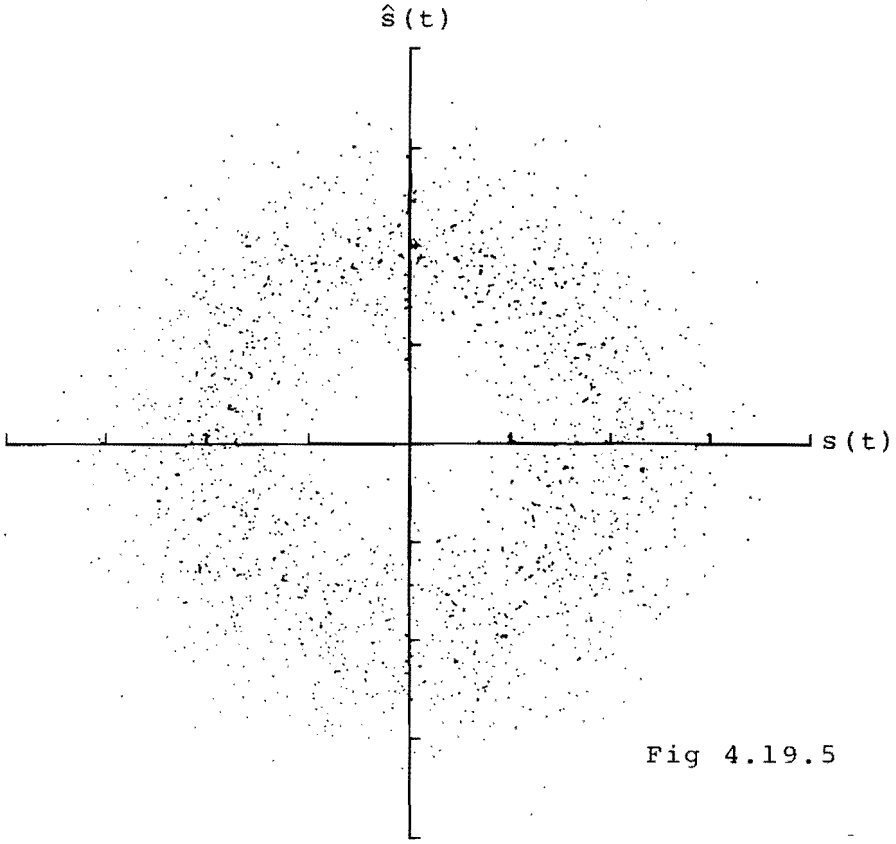


Fig 4.19.5

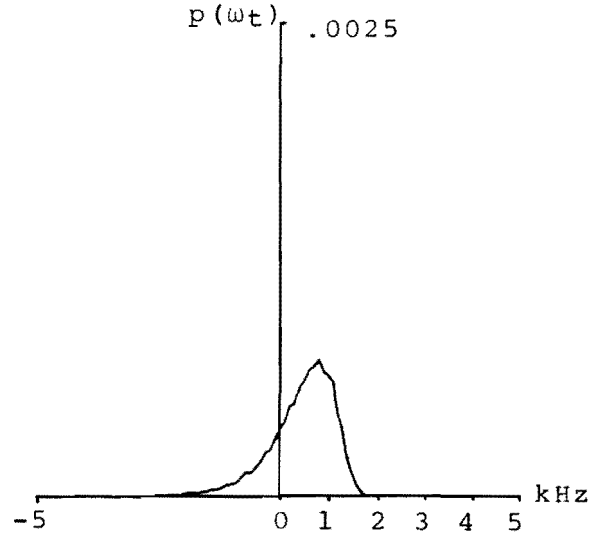


Fig 4.19.6

Figs 4.19.5-4.19.6 Voiced Fricative Model Analysis

the distribution becomes approximately Gaussian with mean $\bar{a}_t=A$. For this reason, instantaneous amplitude rarely approaches zero in figures (4.19.4) and (4.19.5).

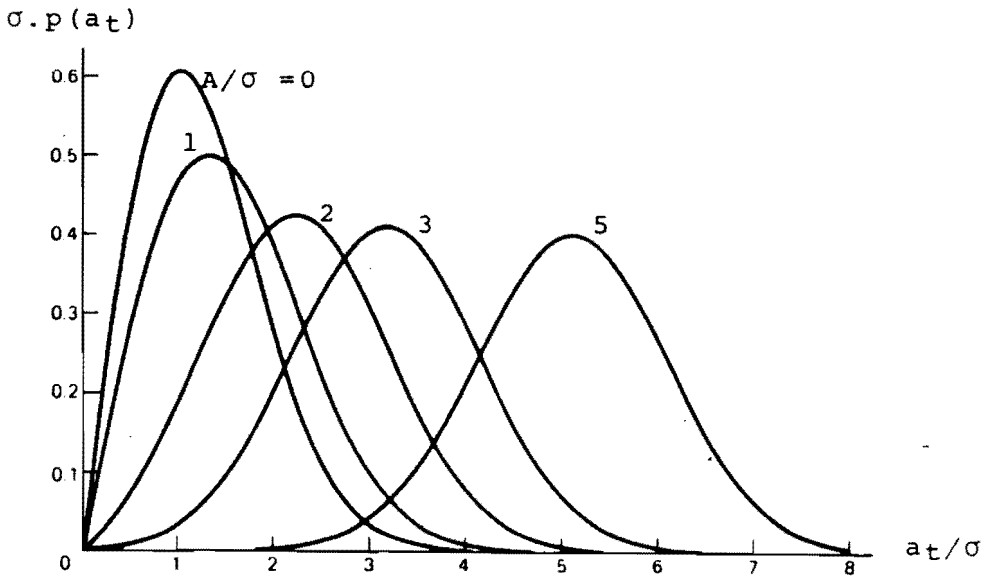


Fig. 4.20 Rician Distribution

The pdf of instantaneous frequency, figure (4.19.6), shows distinct assymetry. Notable features are that $\overline{\omega_i(t)}$ does not correspond to the peak of the distribution and that $p(\omega_i(t)=\omega_m)$ is approximately zero.

For the case of a sinusoid plus rectangularly bandpass filtered noise, Rice (Ref. 114) shows that average instantaneous frequency can take on values between ω_s and ω_m according to

$$\overline{\omega_i} = \omega_m+(\omega_s-\omega_m)\left(1-\exp\left(\frac{-A^2}{2\sigma^2}\right)\right) \quad \ldots (4.19)$$

Research into the use of instantaneous frequency techniques for detection of signals in noise (Ref. 115) has shown that the average instantaneous frequency of a sinusoid plus noise always corresponds to the frequency of the sinusoid when $A/\sigma > 3$.

(4.3.3.2) VOICED FRICATIVE /z/

Figure (4.21.1) is the amplitude spectrum of the voiced fricative /z/ uttered by a male speaker and lowpass filtered to 3,400 Hz. This waveform was not bandpass filtered in an attempt to preserve the low frequency harmonic structure.

The signal bears a strong resemblance to the previous example except that in this case the low frequency periodic signal is made up of about 5 harmonics. Analysis of the first formant of a vowel, Section (4.3.1.1), has shown the form of "basic" instantaneous waveforms associated with such lowpass periodic signals. The "basic" instantaneous amplitude curve of the voiced fricative, $a_p(t)$, will be dominated by major periodic dips due to two or three major analytic signal complex zeros per cycle. Instantaneous frequency, $\omega_p(t)$, will probably display two major rises and one dip per cycle with average instantaneous frequency, $\overline{\omega_p}$, corresponding to the third harmonic.

Unlike the model unvoiced fricative, the existence of "basic" waveforms means that the periodic component of this real signal changes its effective amplitude and frequency with time and we must therefore consider instantaneous values of $a_p(t)/\sigma$ and $\omega_p(t)$. For example, during periods of high instantaneous amplitude ($a_p(t)/\sigma >> 1$) the vector amplitude distribution will be approximately Rician, but during a major dip ($a_p(t)/\sigma << 1$), the amplitude distribution will approach the Rayleigh.

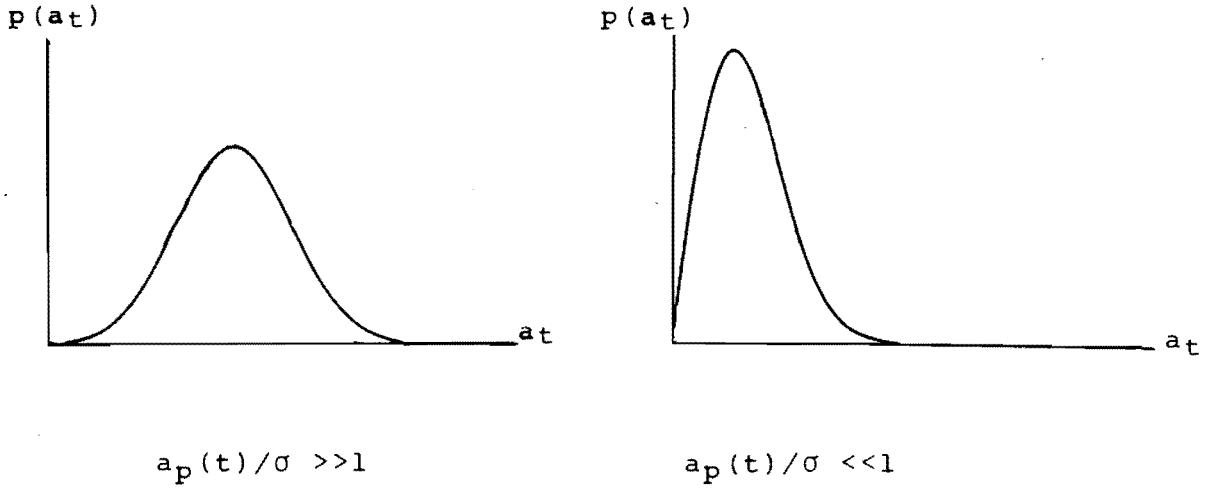


Fig. 4.22 Possible Vector Amplitude Distributions

The pdf of instantaneous frequency will also change during the cycle of the periodic signal. When ap(t)/σ <<1, the pdf of instantaneous frequency resembles that obtained for the unvoiced fricative /s/ (as both /z/ and /s/ employ the same point of vocal tract constriction). When $a_p(t)/\sigma >> 1$, however, the pdf will resemble that of the sinusoid plus noise model, figure (4.19.6).

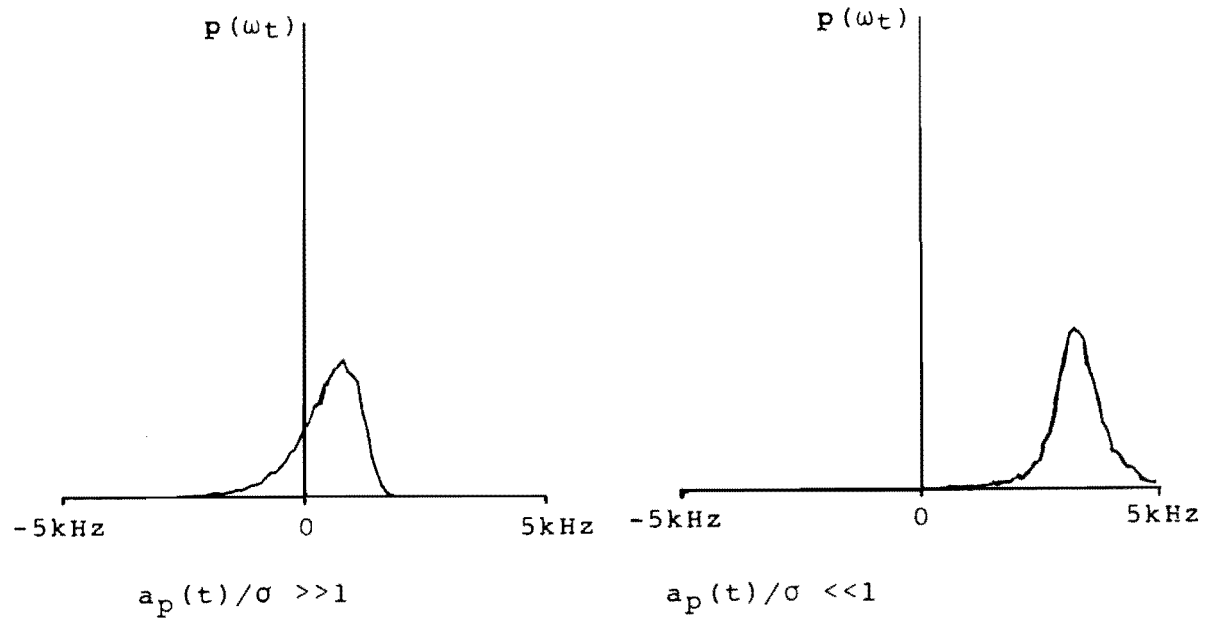


Fig. 4.23 Possible Instantaneous Frequency Distributions

Figure (4.21.2) is the waveform /z/ over a 35 ms period and the corresponding instantaneous waveforms are figures (4.21.3) and (4.21.4). As expected, instantaneous frequency is centered around 300 Hz during the periods of high instantaneous amplitude, but when instantaneous amplitude is low, instantaneous frequency appears to vary wildly. Comparison of figures (4.21.2) and (4.21.4) reveals that the time waveform is noise like during periods of low average instantaneous amplitude. Over the first two cycles, however, the noise component during the period of low a(t) is of insufficient amplitude to cause zero crossings of s(t). This effectively reduces the number of possible UHP analytic signal complex zeros during the period of low instantaneous amplitude and prevents the instantaneous frequency pdf from assuming its symmetrical shape around the short time average instantaneous frequency, $\omega_m (a_p(t)/\sigma << 1)$.

During the third cycle of /z/ in figure (4.21.2), the noise amplitude has increased. Correspondingly, the third cycle of instantaneous frequency exhibits the expected short term average instantaneous frequency shift.

The average instantaneous frequency of the signal for any instantaneous value of $a_p(t)/\sigma$ may be roughly determined from Rice's equation (4.19), substituting $\omega_p(t)$ for ω_s. (Ref. 116). As wild variations of $\omega_p(t)$ can only occur at very low values of $a_p(t)$, these should have little effect on $\overline{\omega_i}$.

(4.3.4) WORD ANALYSIS

In order to rationalise computer storage space and processing times, the sampling rate has been reduced to 10,000 samples per second for all word analyses.

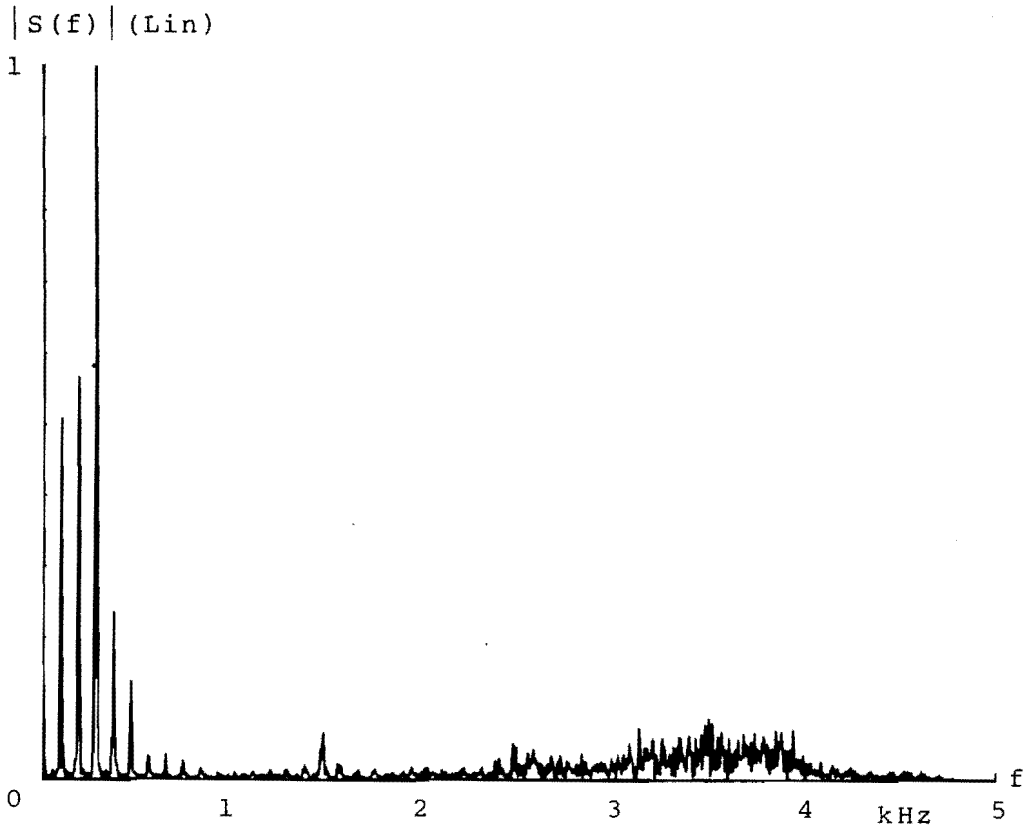


Fig 4.21.1

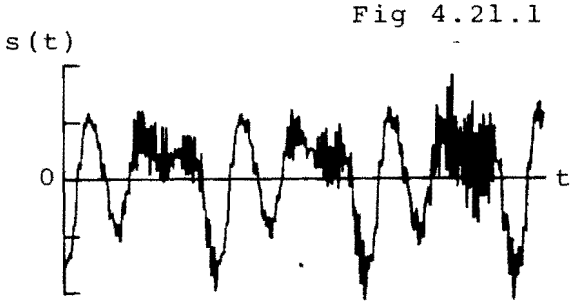


Fig 4.21.2

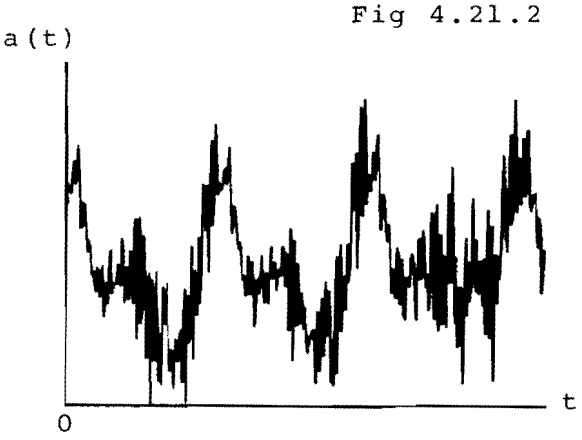


Fig 4.21.4

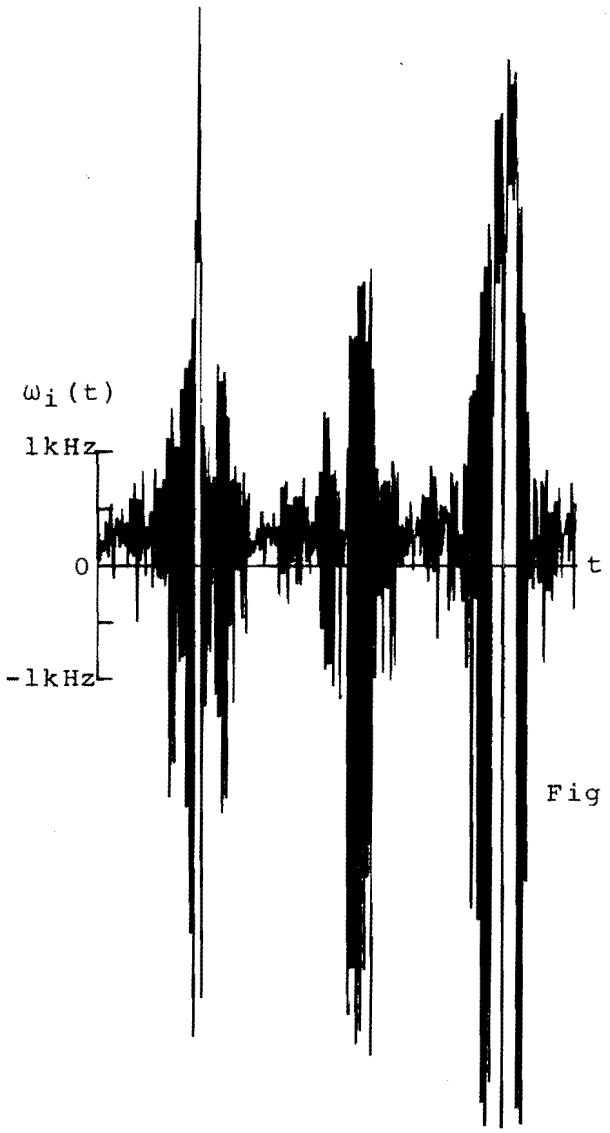


Fig 4.21.3

Figs 4.21.1-4.21.4

Analysis of /z/

(4.3.4.1) "HELLO"

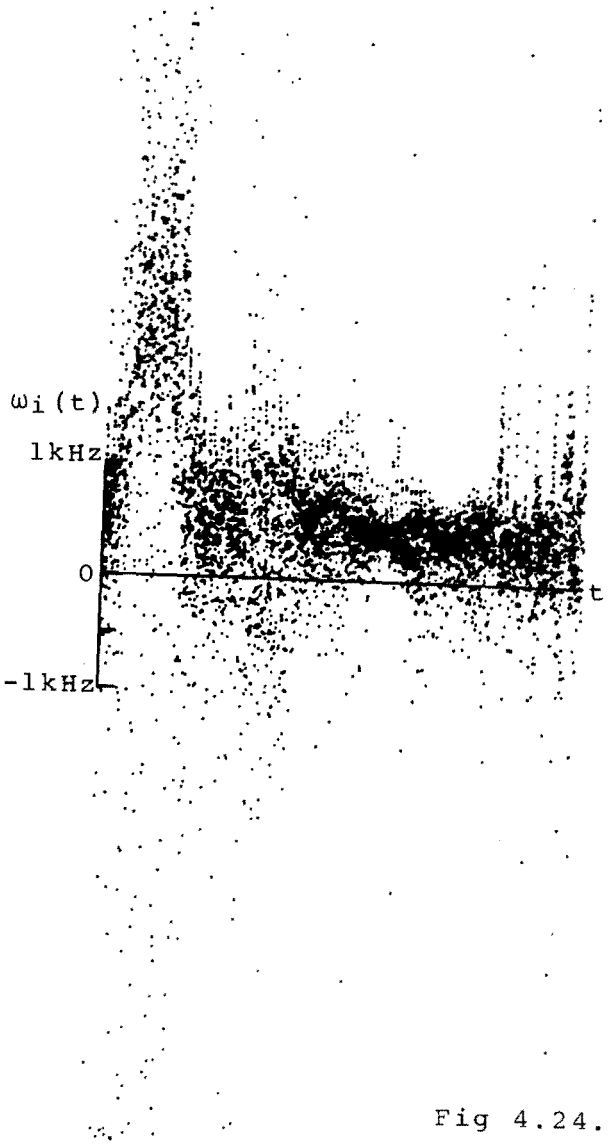
Figures (4.24.1) and (4.24.2) are an instantaneous amplitude and frequency analysis of the word "hello" uttered by a male speaker. Only the instantaneous frequency sample values are plotted in figure (4.24.2) as this gives the picture a less cluttered appearance. The instantaneous parameter analysis is to be viewed in conjunction with figure (4.25) which is the more familiar Fourier analysis of the same word.

As the preceeding investigations have been into the characteristics of individual phonemes, the two types of word analysis are most easily compared on a phoneme by phoneme basis.

The whisper /h/ is expected to be of very low amplitude, and this is confirmed by both the Fourier analysis and instantaneous amplitude plots. Instantaneous frequency, however exhibits the high average value expected for an unvoiced phoneme and appears to accurately mark the phonemes onset and end times. Figure (4.26) is the pdf of the instantaneous frequency during /h/ and this shows the mean frequency to be $\omega_m = 2\pi x 2,300$ radians per second. The pdf of the idle channel, figure (4.27), exhibits a much lower average instantaneous frequency at approximately $\omega_m = 2\pi x 900$ radians per second.

Onset of the vowel /ɛ/ is, indicated by the appearance of formant structure in the Fourier analysis and the start of large periodic instantaneous amplitude fluctuations. At the point of transition from /h/ to /ɛ/, the average instantaneous frequency drops to approximately 700 Hz, resulting in the pdf of figure (4.28).

The principally low frequency periodic energy of the semivowel /ℓ/ is visible in terms of the high amplitude, low frequency harmonics in the Fourier analysis and in terms of the dense clustering of instantaneous frequency



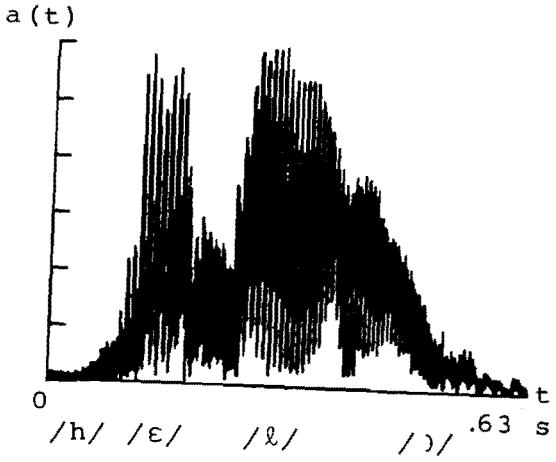


Fig 4.24.1

Fig 4.24.2

Fig 4.24 Instantaneous Parameter Analysis of "HELLO"

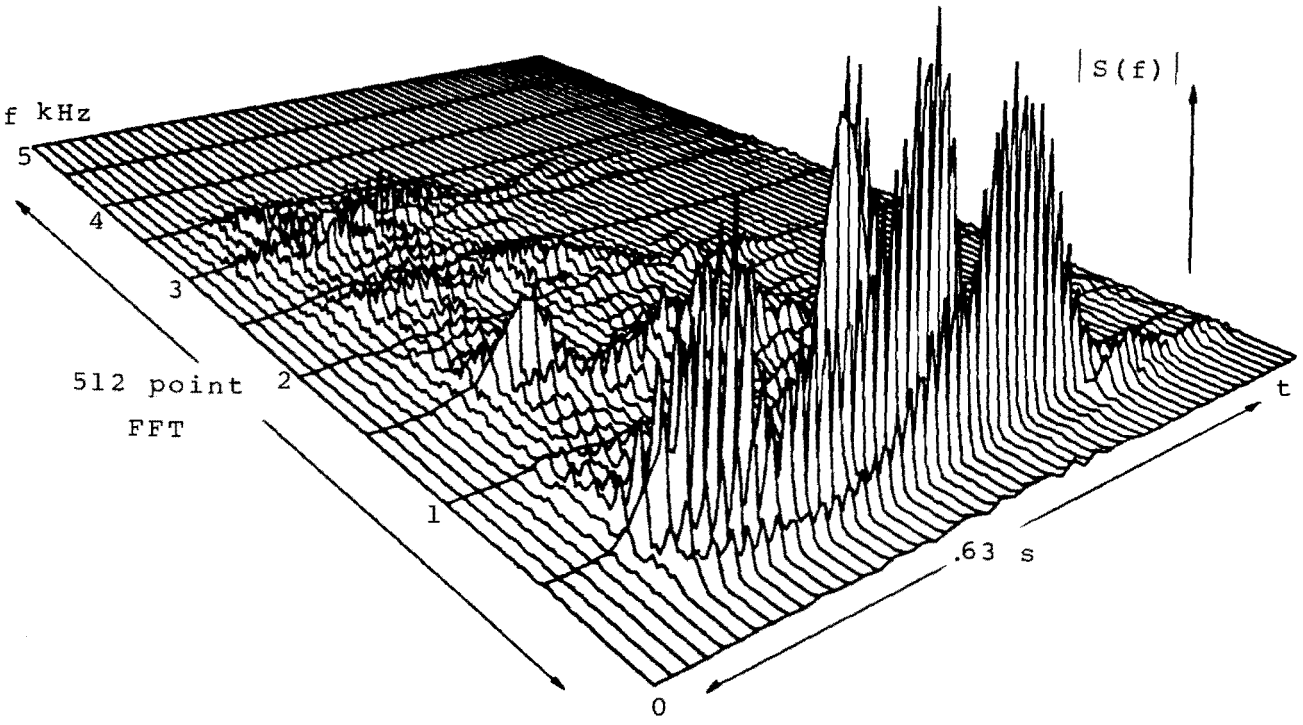


Fig 4.25 Fourier Analysis of "HELLO"

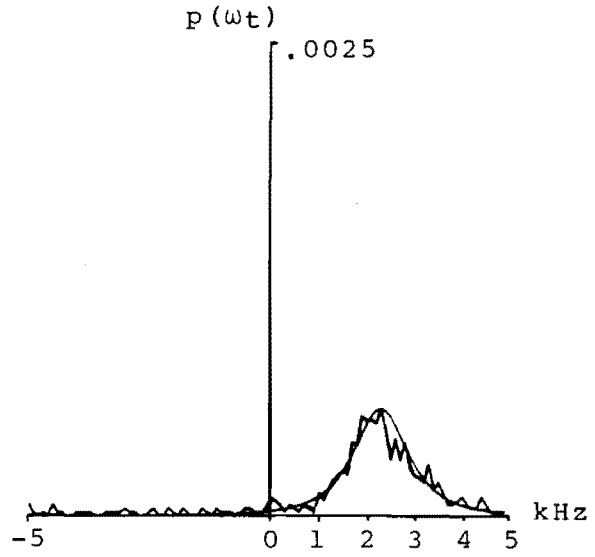


Fig 4.26 pdf of /h/

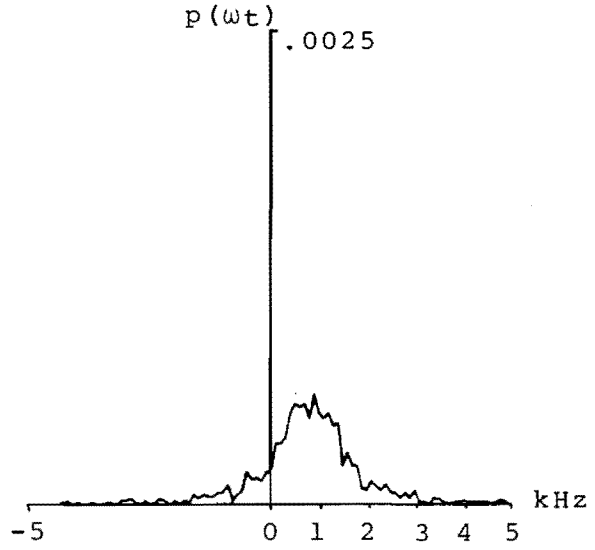


Fig 4.27 pdf of Idle Channel

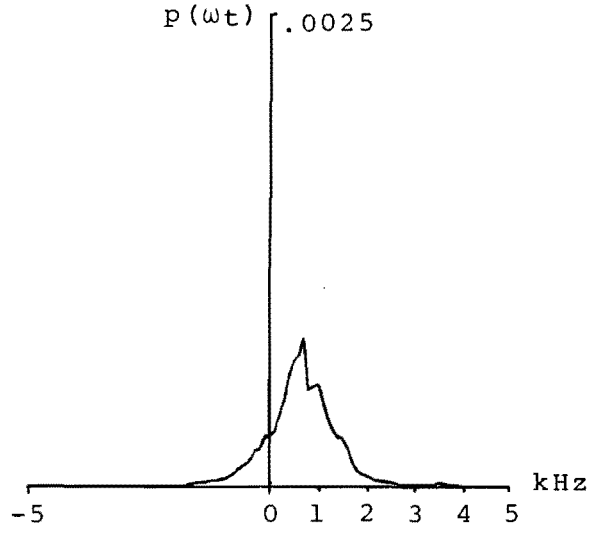


Fig 4.28 pdf of /ε/

around approximately 400 Hz. The relatively high amplitude of the phoneme is clearly displayed by instantaneous amplitude.

As /ℓ/ is decaying, it "glides" into the vowel /ɔ/. The vowel is of low amplitude and short duration and is therefore difficult to detect in either the Fourier or instantaneous parameter analysis.

The most striking feature of instantaneous frequency analysis of the word "hello" is the transition from unvoiced /h/ to voiced /ɛ/. Figure (4.29.1) is the time waveform during this transition period and figures (4.29.2) and (4.29.3) are the instantaneous waveforms, although over a longer time interval. An indication of the average instantaneous frequency at each instantaneous frequency sample, $\omega_i(kT)$, in the transition can be obtained by forming a pdf from the n previous and n following instantaneous frequency samples. (i.e. $\omega_i\{(k-n)T\}$ to $\omega_i\{(k+n)T\}$) This creates a "dynamic" pdf from 2n+1 samples whose peak, in the use of unvoiced fricatives, may be deemed the average instantaneous frequency at the time kT.

Figure (4.30.1) is average instantaneous frequency calculated from a dynamic pdf of 11 samples length (n=5) for the instantaneous frequency data in figure (4.29.2). The estimated average varies wildly during both the unvoiced and voiced phoneme, suggesting that n may be too low. Increasing the dynamic pdf to 101 points (n=50) produces the average instantaneous frequency estimate, figure (4.30.2). In this case, the average is better defined and the onset and end of the phoneme /h/ are clearly marked.

From the point of view of real time operation, using a dynamic pdf to mark transitions between phonemes requires

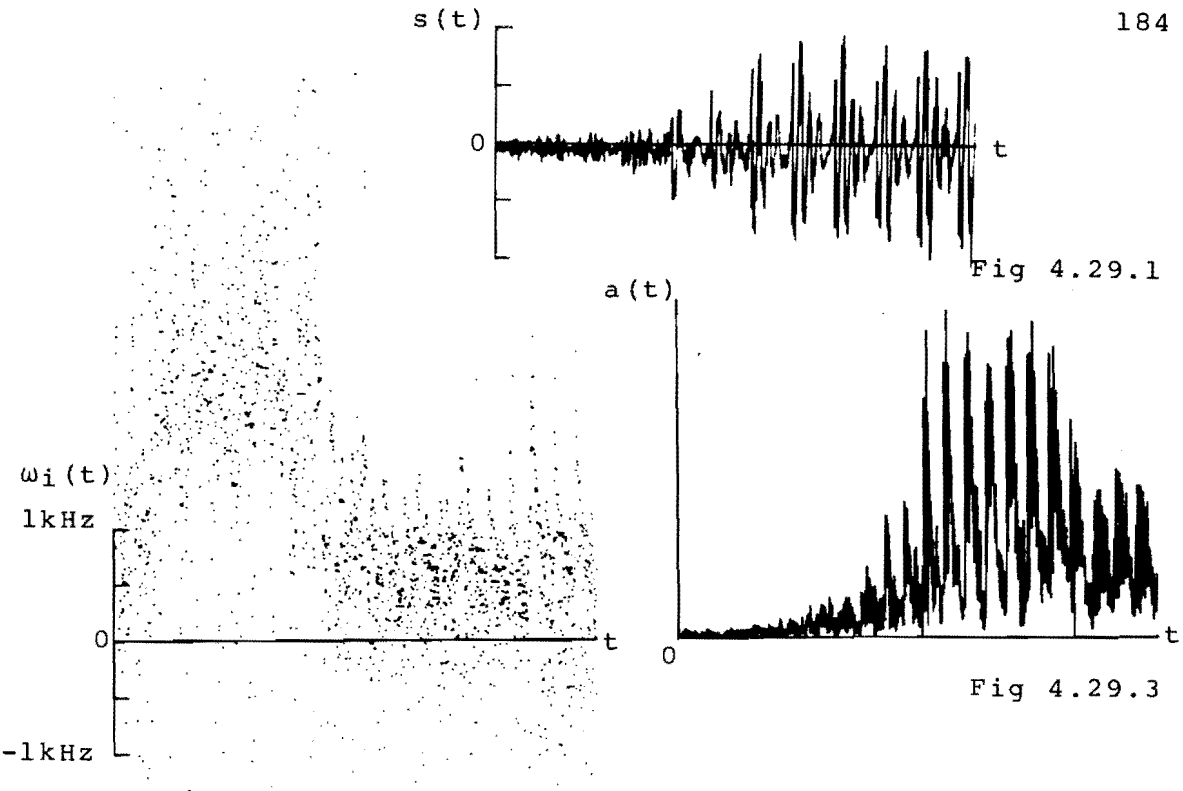


Fig 4.29.1

Fig 4.29.3

Fig 4.29.2

Fig 4.29 Analysis of /h/→/ɛ/ Transition

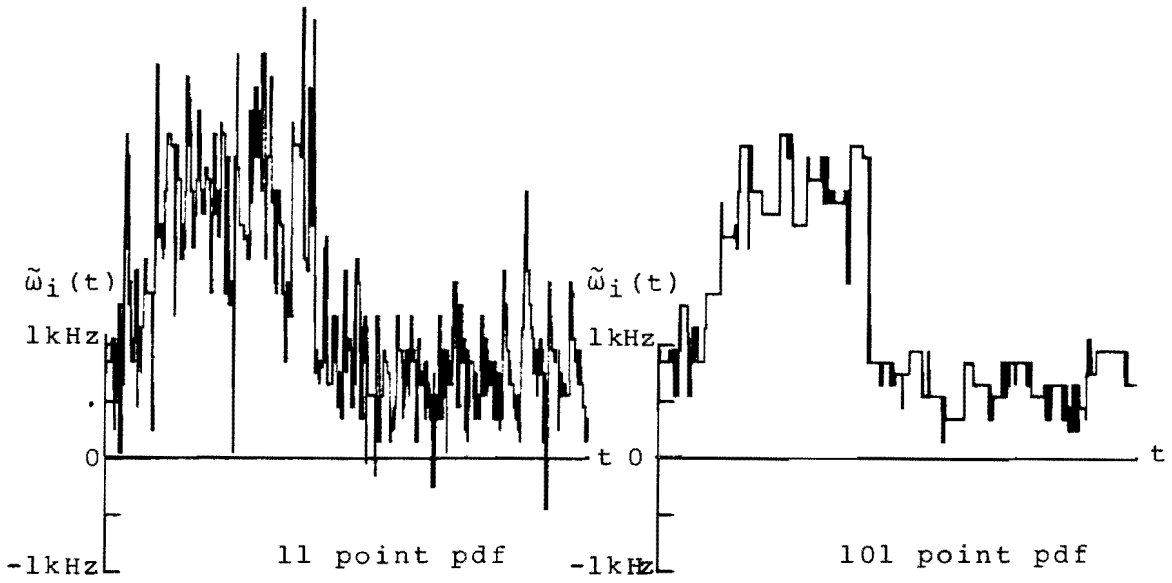


Fig 4.30.1

Fig 4.30.2

Fig 4.30 Dynamic pdf Analysis of /h/→/ɛ/ Transition

looking forward for n sampling periods. This could be achieved by delaying the waveform by n samples, which corresponds to a time of 5 ms in the case of figure (4.30.2).

(4.3.4.2) "SET"

Figures (4.31.1) and (4.31.2) are instantaneous amplitude and frequency analyses of the word "set". Once again, instantaneous amplitude displays high amplitude periodic fluctuations during the voiced phoneme and instantaneous frequency marks the transition between voiced and unvoiced phonemes by abrupt changes of its pdf. Both instantaneous amplitude and frequency indicate the rapid onset of the unvoiced plosive /t/.

The time waveform over the transition /s/ to /ɛ/ is shown in figure (4.32.1) along with the corresponding instantaneous functions figures (4.32.2) and (4.32.3). Applying the dynamic pdf technique of determing average instantaneous frequency to the data in figure (4.32.2). results in figure (4.33.1), the estimated average for a pdf of 51 samples (n=25) and figure (4.33.2), the estimated average for a pdf of 101 samples (n=50). Although the phoneme transition is indicated in both estimations, the point is more precisely marked in figure (4.33.2).

To illustrate the type of data from which figure (4.33.2) is constructed, two pdfs of length 101 points were generated from different points in a sustained version of the phoneme /s/, uttered out of context. These are figures (4.34.1) and (4.34.2). It can be seen that the distributions are not smooth and each indicates a different average value. The pdfs could be smoothed by increasing the number of instantaneous frequency samples used in their formation, but this would cause greater time delays in a real time implementation of this technique and could cause problems in estimating the average instantaneous frequency of short duration phonemes.

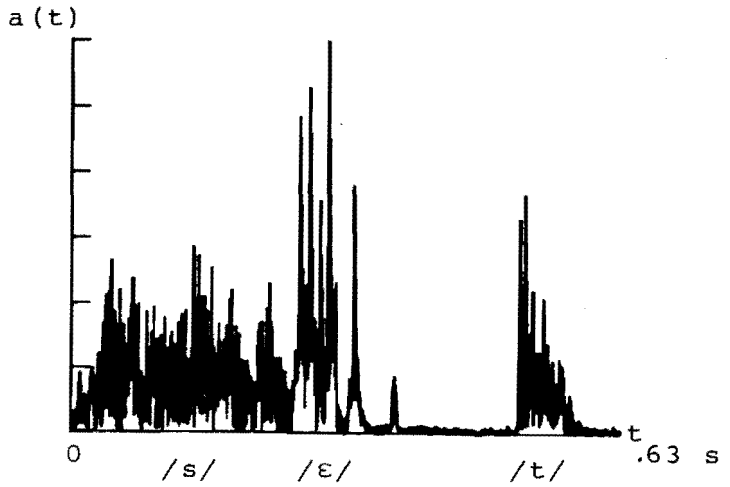


Fig 4.31.1

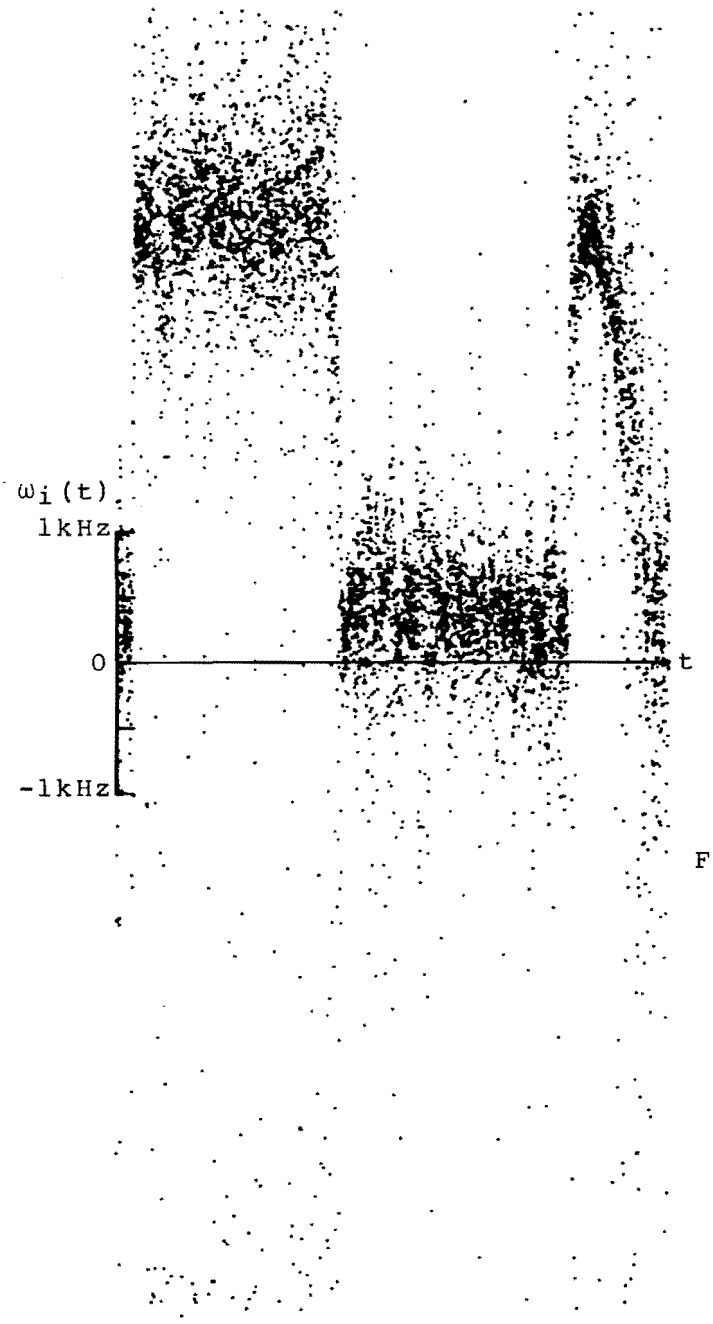


Fig 4.31.2

Fig 4.31 Instantaneous Parameter Analysis of "SET"

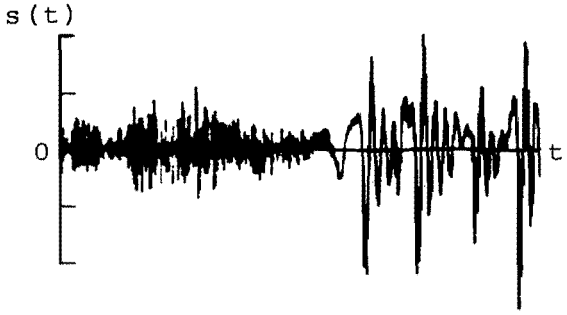


Fig 4.32.1

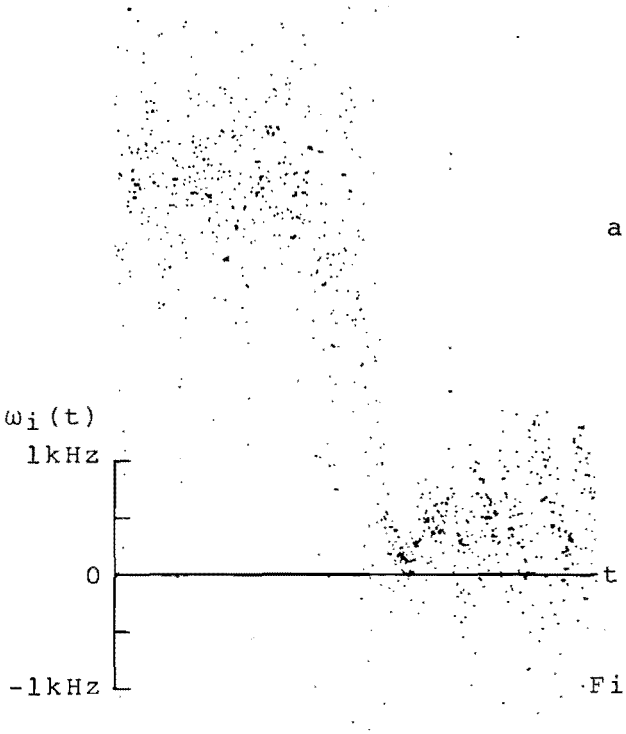


Fig 4.32.2

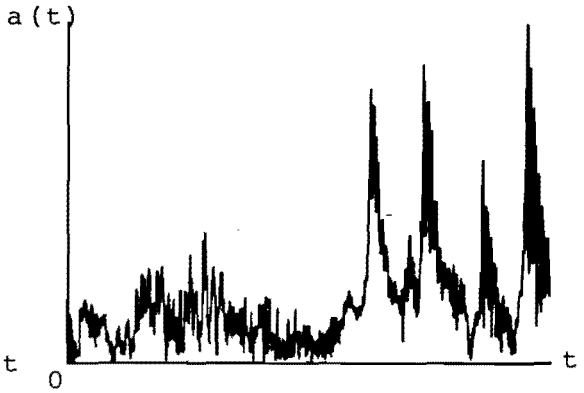


Fig 4.32.3

Fig 4.32 Analysis of /s/→/ɛ/ Transition

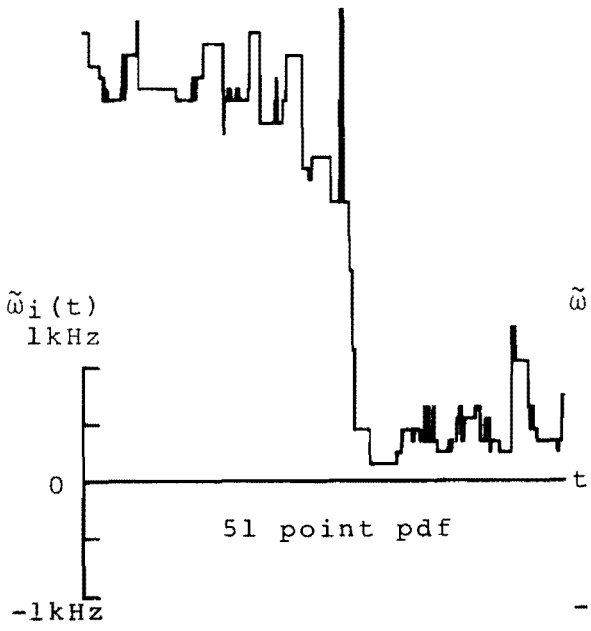


Fig 4.33.1

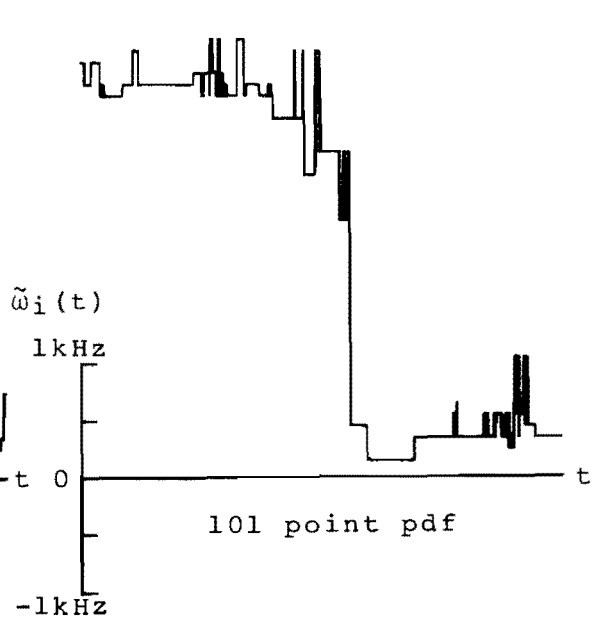


Fig 4.33.2

Fig 4.33 Dynamic pdf Analysis of /s/→/ɛ/ Transition

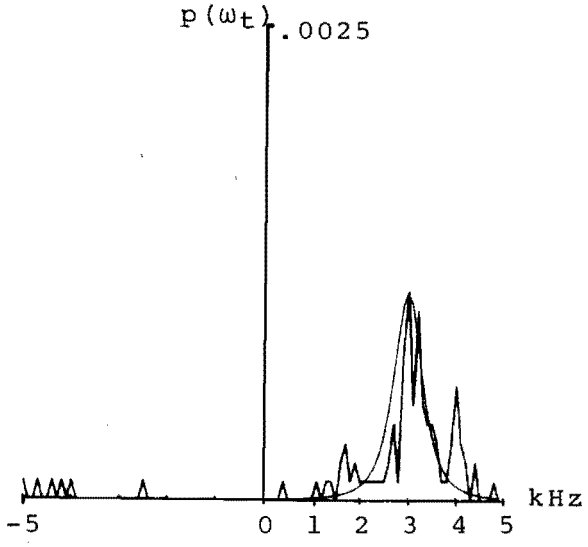


Fig 4.34.1

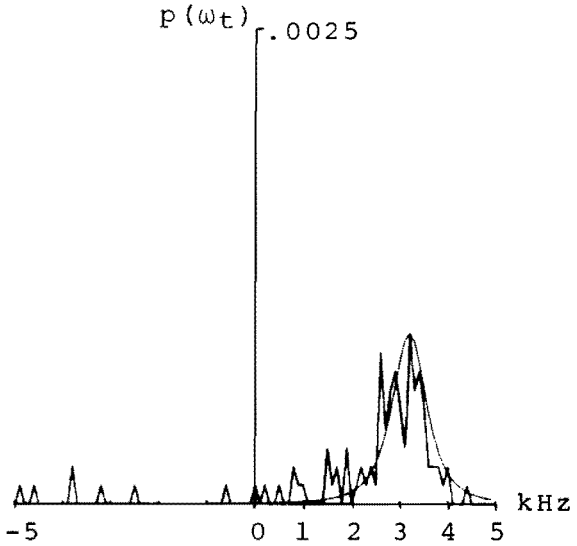


Fig 4.34.2

Fig 4.34 101 Point pdfs from Sustained /s/

The time waveform of the onset and duration of the plosive /t/ is illustrated in figure (4.35.1) and figures (4.35.2) and (4.35.3) are the corresponding instantaneous functions. It appears, from figure (4.35.2), that the pdf of instantaneous frequency is similar to that of other unvoiced fricatives, except that average instantaneous frequency falls over the phoneme duration. This is confirmed by the 101 point dynamic pdf estimation, figure (4.35.4), which also clearly indicates the moment of onset.

(4.3.4.3) "FAST"

The final instantaneous parameter analysis is of the unbandlimited word "fast". Figure (4.36.1) is the instantaneous amplitude, figure (4.36.2) the instantaneous frequency and figure (4.36.3) a 101 point dynamic pdf analysis of the instantaneous frequency data.

The phoneme /f/ is similar to the whisper /h/ in "hello" as it is of low amplitude and is the first phoneme of the word. Accordingly, instantaneous amplitide is very low over its duration, but instantaneous frequency shows a reasonably well defined pdf. The estimated average instantaneous frequency is not constant over /f/, but it accurately indicates the phonemes onset and end.

Unlike the whisper /h/, the vocal tract cannot move from /f/ to the following vowel without a pause. The short gap between /f/ and /æ/ is not noticeable in figures (4.36.1) or (4.36.2), but shows as a period of confused average instantaneous frequency in figure (4.36.3).

The onset of /æ/ is marked by periodic instantaneous amplitude fluctuations and a sudden reduction of average instantaneous frequency. The amplitude of the vowel falls over its duration and the vocal tract eventually moves, without noticeable pause, into the unvoiced fricative /s/.

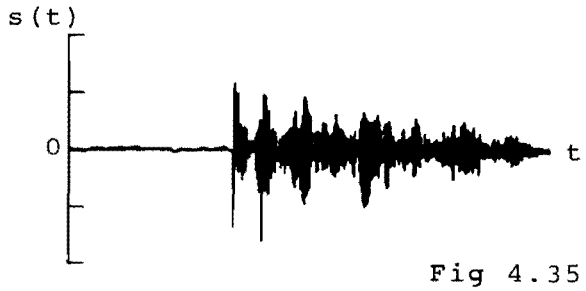
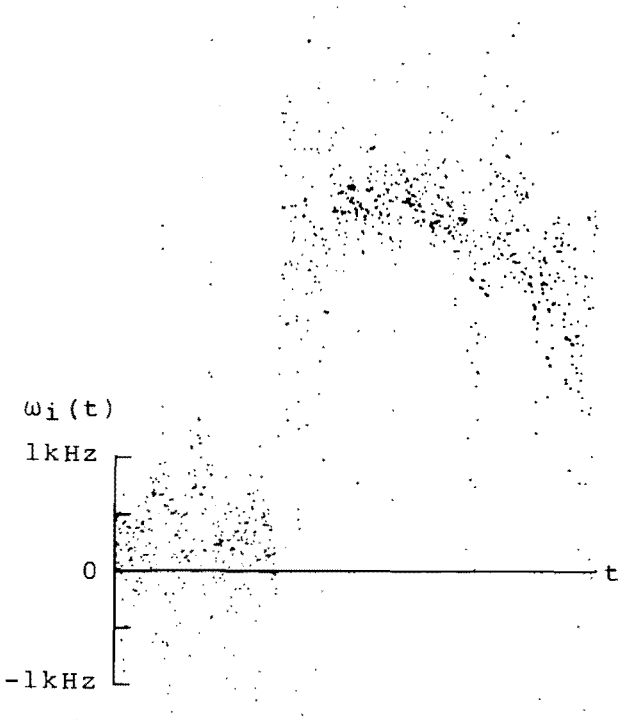


Fig 4.35.1



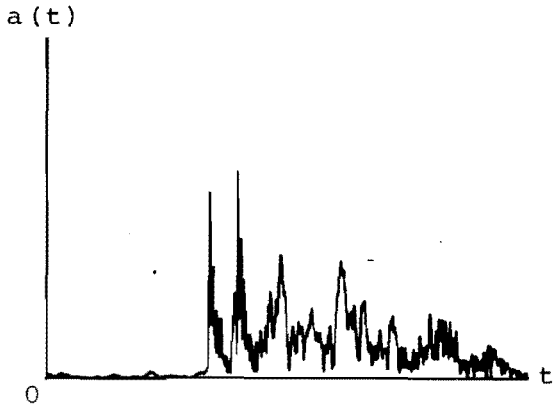


Fig 4.35.3

Fig 4.35.2

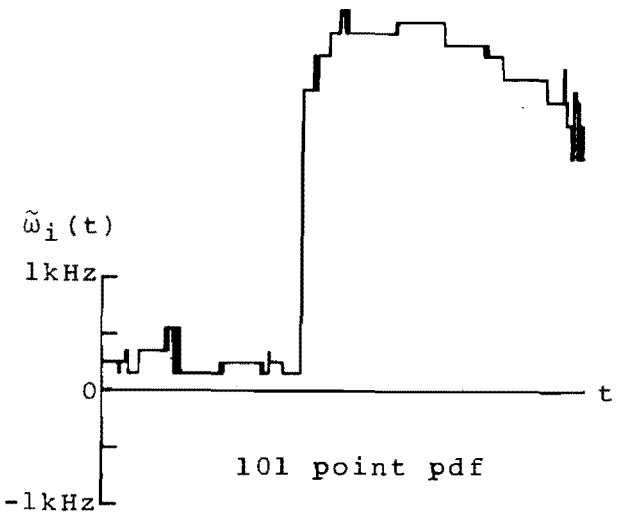


Fig 4.35.4

Fig 4.35 Analysis of Onset and Duration of /t/

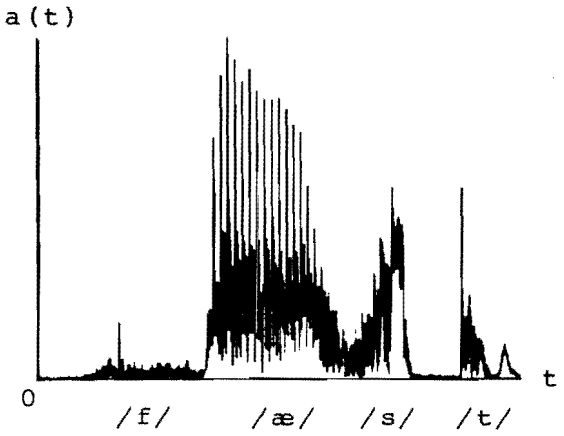


Fig 4.36.1

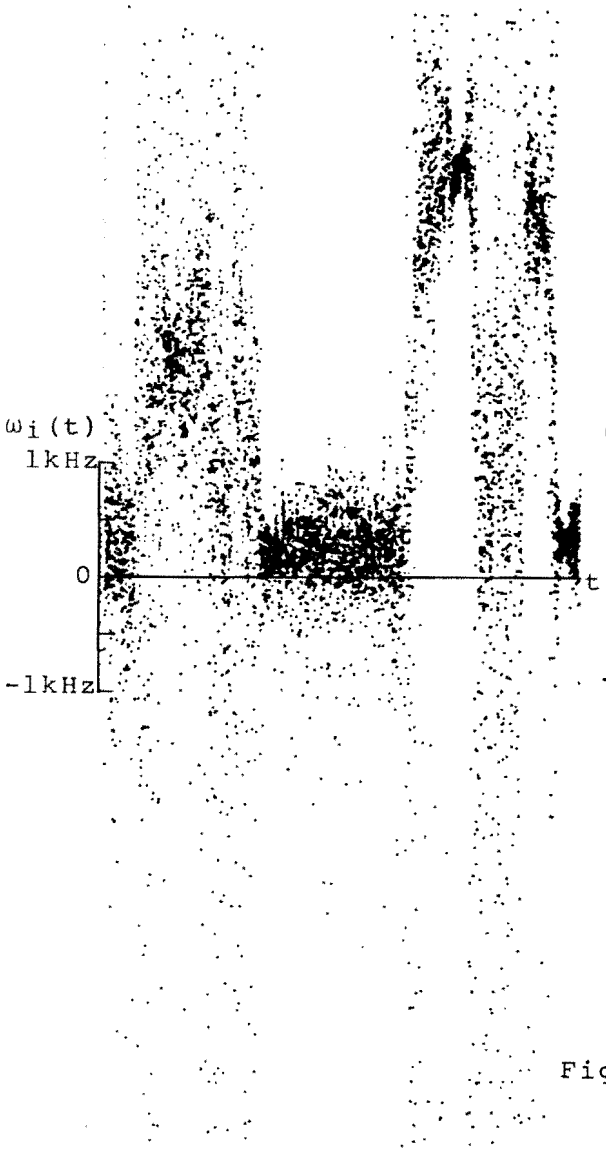


Fig 4.36.2

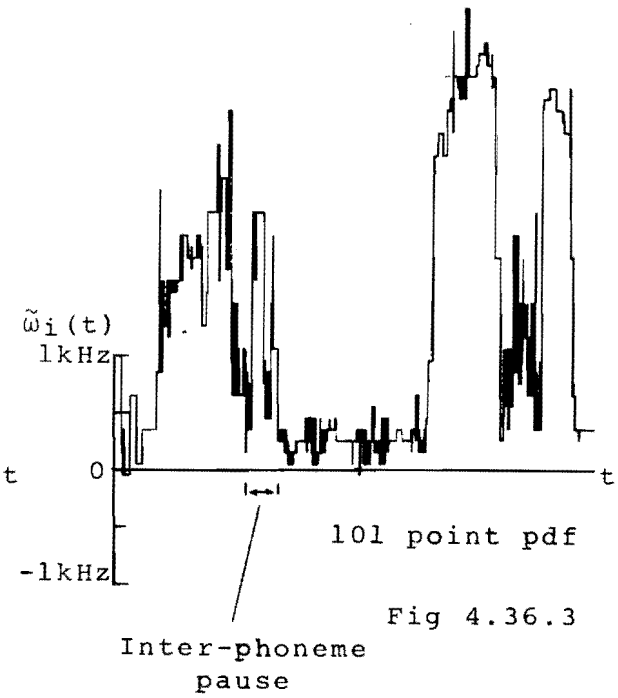


Fig 4.36.3

Fig 4.36 Analysis of "FAST"

Both the instantaneous amplitude and average instantaneous frequency of /s/ rise over its duration until just before the phonemes end, the instantaneous frequency pdf is very narrow. The narrow pdf indicates that /s/ has become a whistle, or nearly a sinusoid at frequency 3,700 Hz.

The gap between /s/ and /t/ is characterised by very low instantaneous amplitude and confused average instantaneous frequency.

The onset of /t/, however, is clearly marked by sudden rises of instantaneous amplitude and average instantaneous frequency, which both fall over the phonemes duration. The end of /t/ is indicated by a sharp fall of average instantaneous frequency and the phoneme is followed by a short period of voicing.

CHAPTER 5

(5.1) INTRODUCTION

This section documents the investigation of signal waveforms reconstructed directly from their instantaneous parameters or from processed instantaneous parameters. The resulting distortions are examined in terms of waveform characteristics, spectra and the intelligibility of processed speech.

A form of bandwidth efficient speech transmission based on reconstruction from narrow bandwidth instantaneous parameters is also investigated.

(5.2) DIRECT RECONSTRUCTION & RESULTING DISTORTIONS

A real waveform may be reconstructed directly from the instantaneous parameters of its associated analytic signal by the relation

$$\tilde{s}(t) = a(t)\cos\{\int\omega_i(t).dt+\phi\} \quad . \ . \ . \ (5.1)$$

Unless special care is taken, the absolute phase of the reconstruction, ϕ, will not correspond to the phase of the original signal. The resulting phase distortion is undetectable in speech listening tests as it is not frequency selective and does not alter the signals amplitude spectrum.

Reconstructions by equation (5.1) can result in significant waveform distortion, if the instantaneous frequency, $\omega_i(t)$, has been calculated or stored with insufficient sampling rate or dynamic range. Distortion of instantaneous frequency fluctuations can change the area enclosed by the instantaneous frequency curve, thus

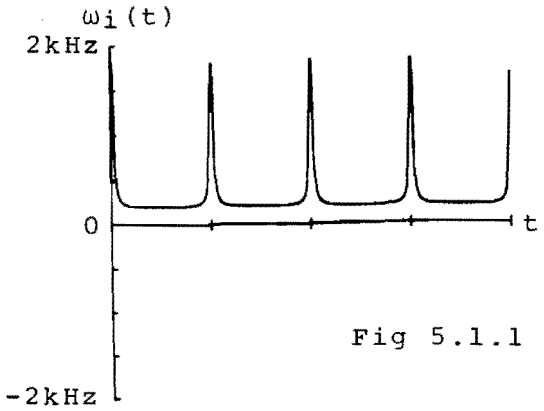


Fig 5.1.1

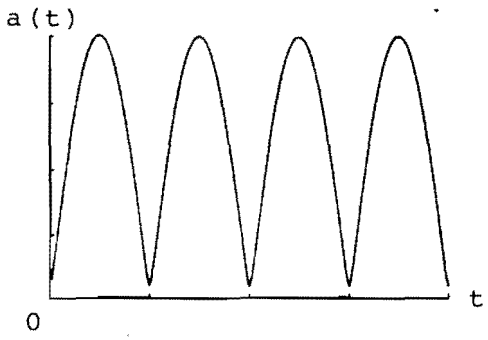


Fig 5.1.2

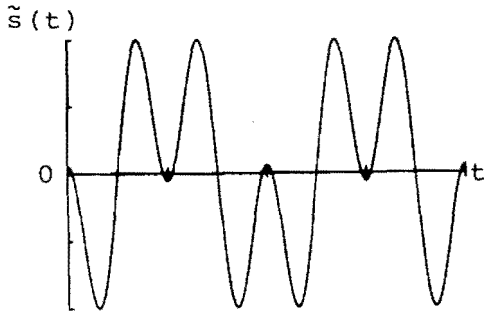


Fig 5.1.3

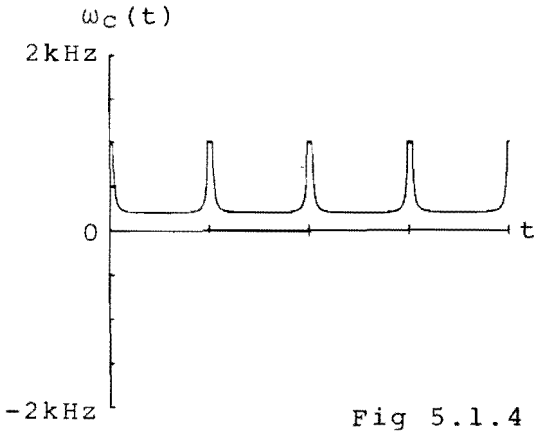


Fig 5.1.4

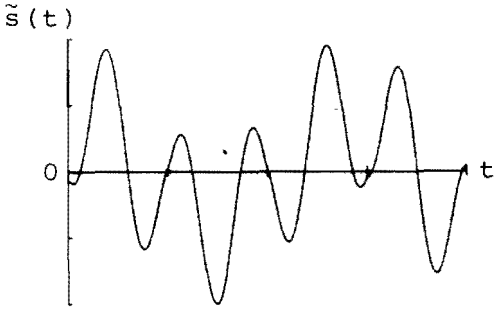


Fig 5.1.5

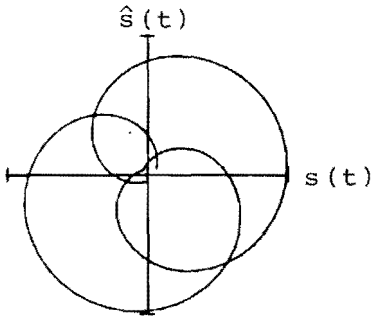


Fig 5.1.6

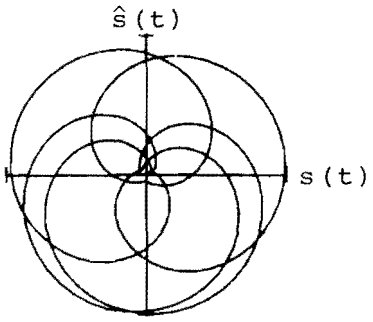


Fig 5.1.7

Fig 5.1 Undistorted and Distorted Reconstruction

inducing phase modulation of the reconstructed signal.

Accurate and distorted reconstruction of the signal s(t),

$$s(t) = \cos(2\pi.100t) - 1 \cdot 1\cos(2\pi.300t) \quad . \; . \; . \; (5.2)$$

is illustrated in figure (5.1). Figures (5.1.1) and (5.1.2) are two cycles of the instantaneous parameters of s(t). Direct reconstruction by equation (5.1) (ϕ=0) yields the undistorted waveform, figure (5.1.3). This is the case B=1.1 from example (2.5.1).

Figure (5.1.4) is the instantaneous frequency waveform dynamic range limited to ± 1,000 Hz. Reconstructing with the original instantaneous amplitude waveform and dynamic range limited instantaneous frequency yields the distorted signal, figure (5.1.5). Apart from the expected phase distortion, the time waveform exhibits slight fluctuations at the points of minimum instantaneous amplitude.

The form and degree of phase modulation is revealed by the vector plot of one cycle, figure (5.1.6). Over one period of the fundamental frequency, the vector rotates only 5.75π radians, indicating that each clipped instantaneous frequency spike results in a loss of approximately 3π/8 radians of phase shift. Figure (5.1.7) is the vector plot over the full two cycles of the fundamental. The vector plot of an undistorted reconstruction would be similar to that of figure (2.19).

The effects of instantaneous frequency dynamic range limitation on a reconstructed voiced phoneme are illustrated in figures (5.2) to (5.4) for the bandpass vowel /ɛ/. Figures (5.2.1) and (5.2.2) are two cycles of the vowels instantaneous amplitude and frequency functions, and direct reconstruction from these waveforms yields the undistorted

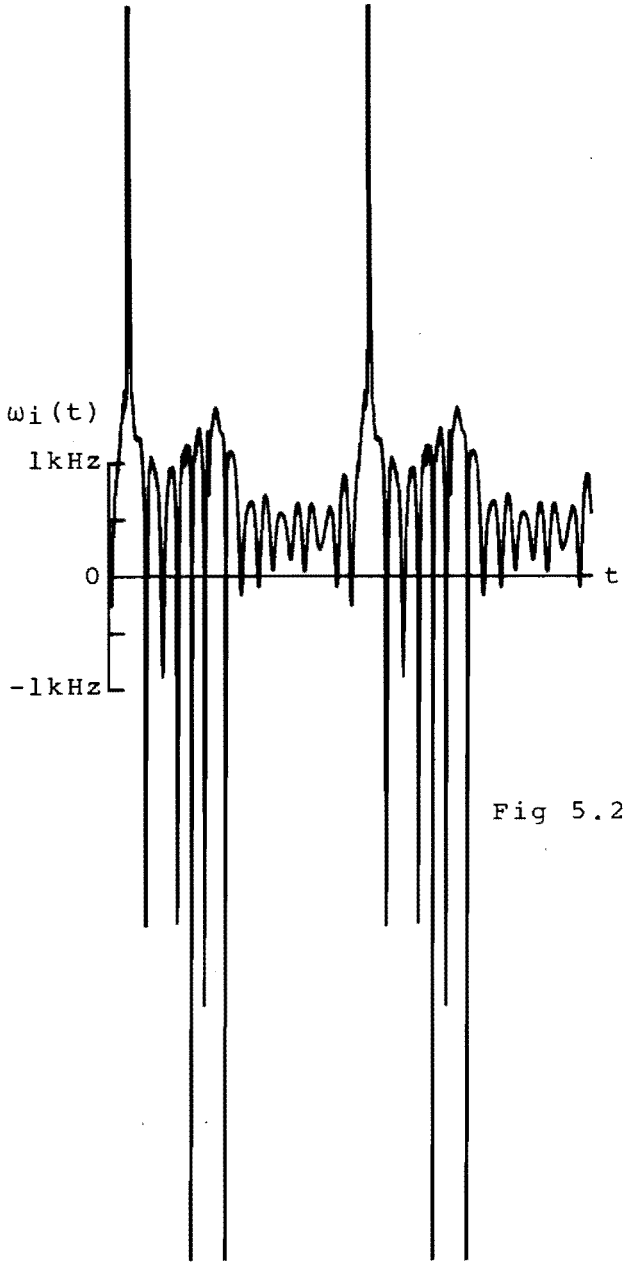


Fig 5.2.1

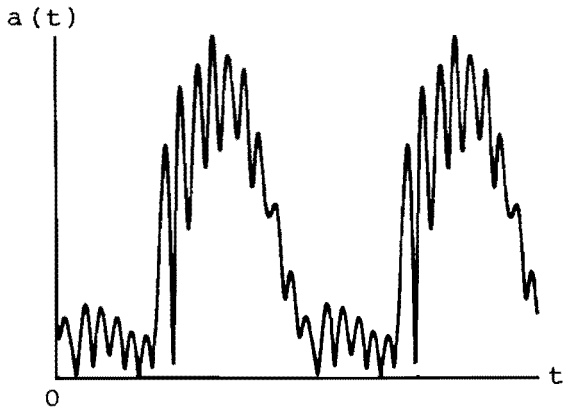


Fig 5.2.2

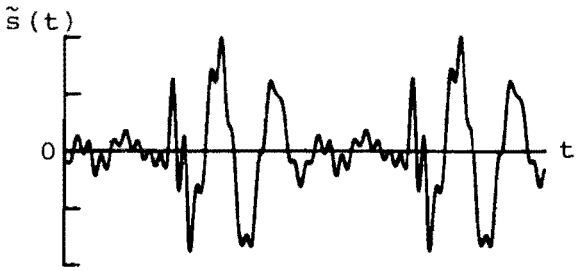


Fig 5.2.3

Figs 5.2.1-5.2.3 Undistorted Reconstruction of /ɛ/

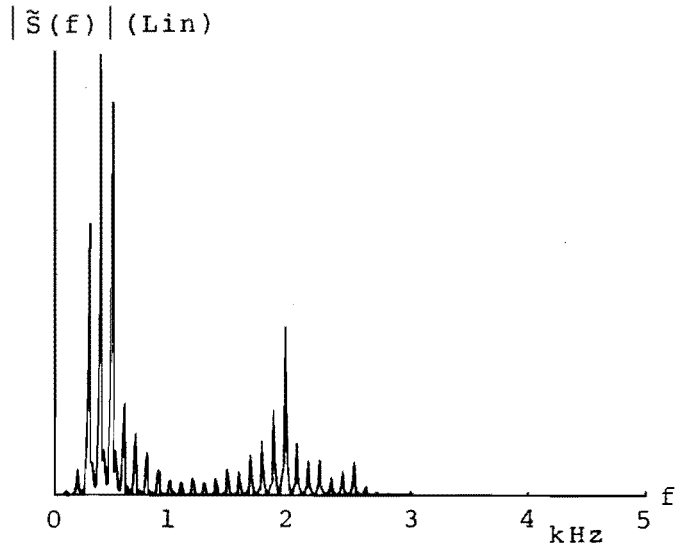


Fig 5.2.4

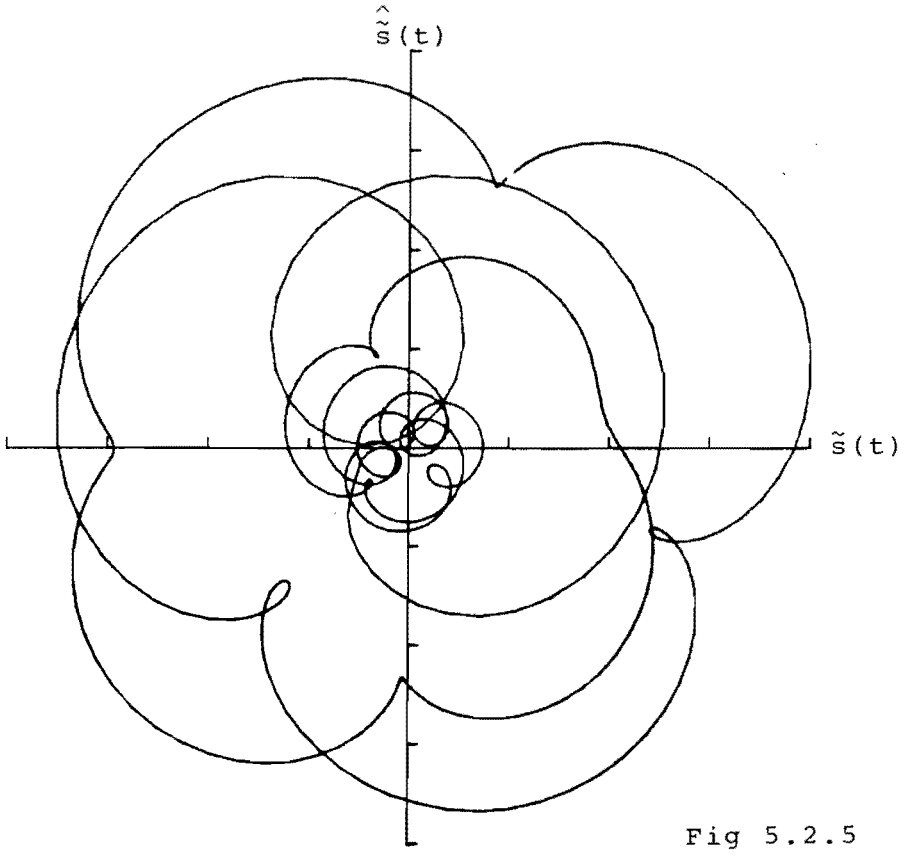


Fig 5.2.5

Figs 5.2.4-5.2.5 Undistorted Reconstruction of /ɛ/

signal, amplitude spectrum and vector locus, figures (5.2.3), (5.2.4) and (5.2.5). The vector rotates exactly 12π radians over one cycle of the fundamental frequency.

Setting a dynamic range limit of ± 3,000 Hz causes instantaneous frequency to clip six times per cycle, figure (5.3.1). The signal reconstructed from this distorted instantaneous frequency curve and the original instantaneous amplitude, figure (5.3.2), displays some phase modulation and its amplitude spectrum, figure (5.3.4) is also distorted. The vector locus over one cycle of the fundamental frequency, figure (5.3.4) now shows a total rotation of approximately 12.3π radians.

The additional 0.3π radians per cycle of the fundamental frequency results from more enclosed area being removed by clipping instantaneous frequency at -3,000 Hz than at +3,000 Hz. Assuming a fundamental frequency of 100 Hz, the additional phase change per cycle shifts the entire vowel spectrum up in frequency by $(0.3\pi/2\pi)\times100=15$ Hz. The frequency of the m^{th} spectral line has therefore become (mx100+15) Hz. The frequency shift is measurable on a full scale version of figure (5.3.3).

Reducing the instantaneous frequency dynamic range to ± 1,500 Hz, figure (5.4.1), results in more severe clipping and the changing phase is again visible on the time waveform, figure (5.4.2). The amplitude spectrum, figure (5.4.3), now displays considerable distortion, but the first and second formants are still clearly visible. The vector plot shows that the waveform advances by an extra π/2 radians during one period of 100 Hz. This causes an amplitude spectrum shift of +25 Hz, and the effect is measurable on a full scale version of figure (5.4.3).

Spectral distortion introduced by reconstruction from dynamic range limited instantaneous frequency has two major

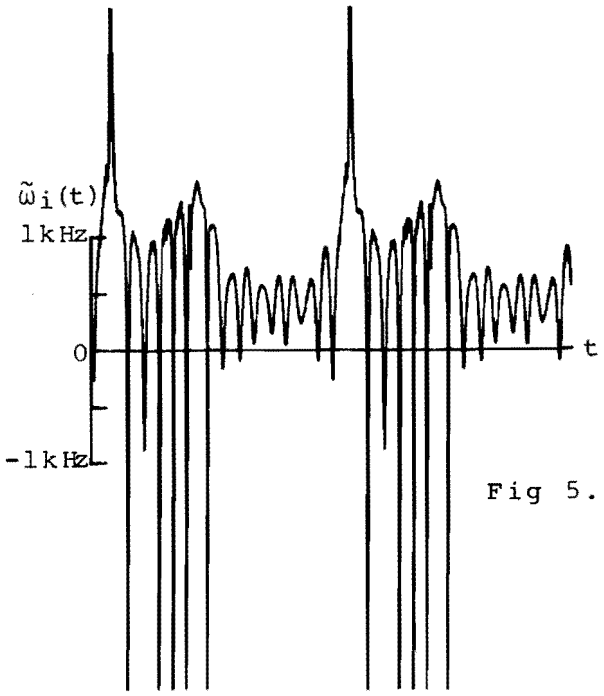


Fig 5.3.1

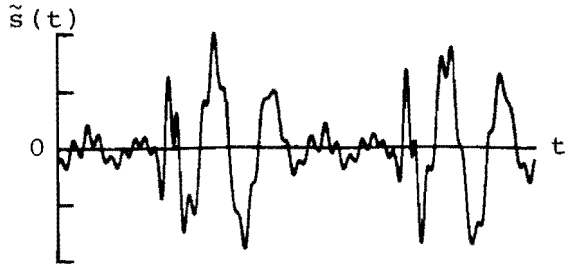


Fig 5.3.2

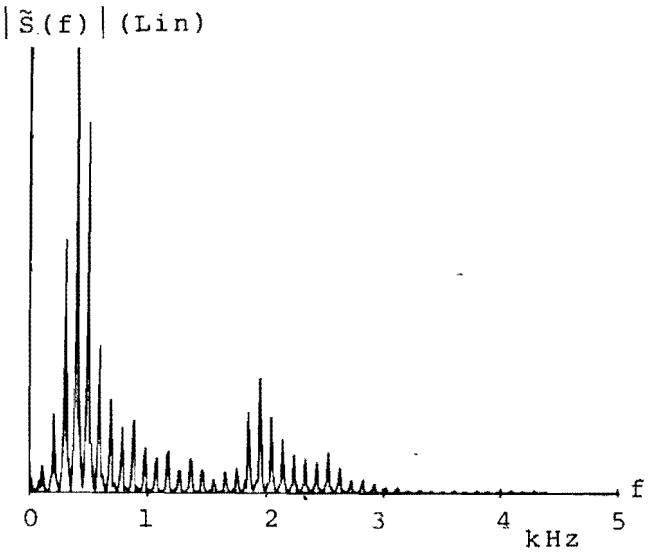


Fig 5.3.3

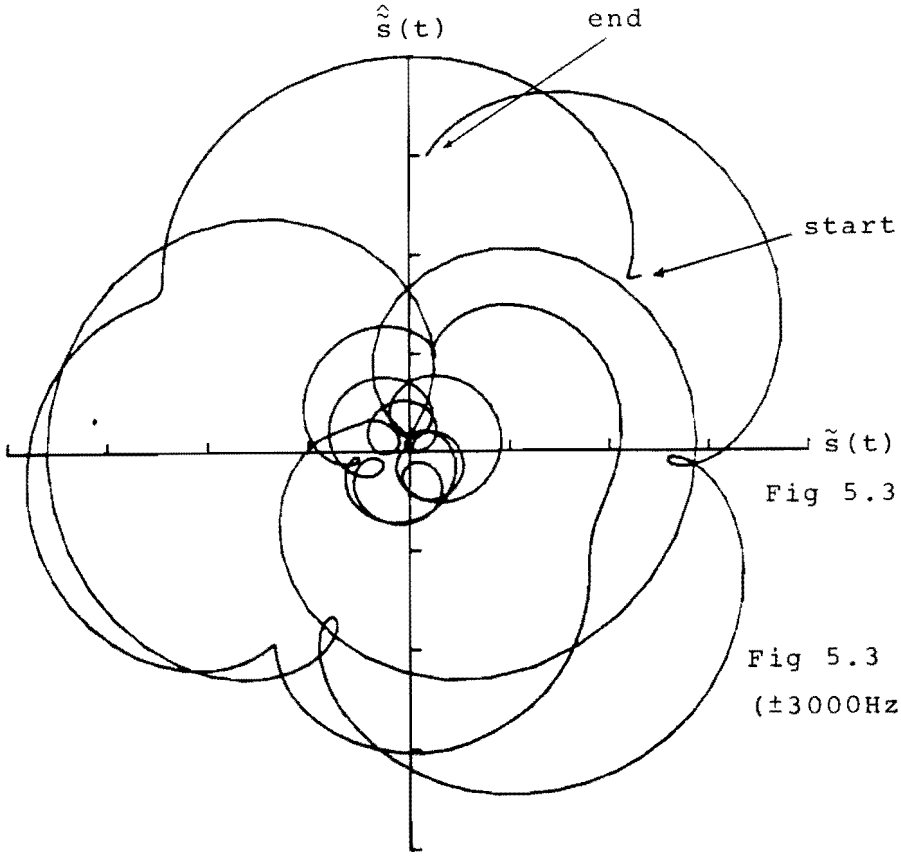


Fig 5.3.4

Fig 5.3 Clipped ωi(t) (±3000Hz) Reconstruction

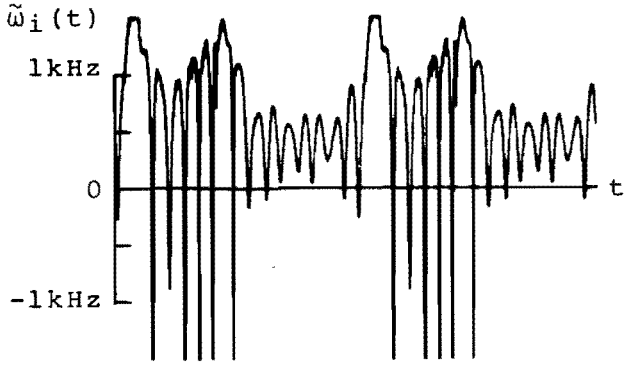


Fig 5.4.1

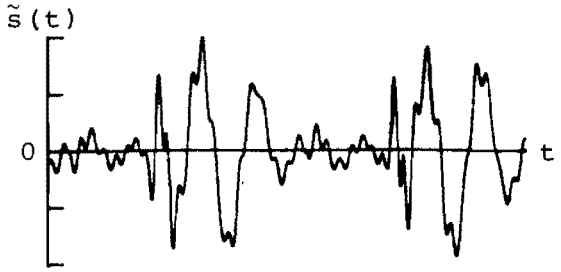


Fig 5.4.2

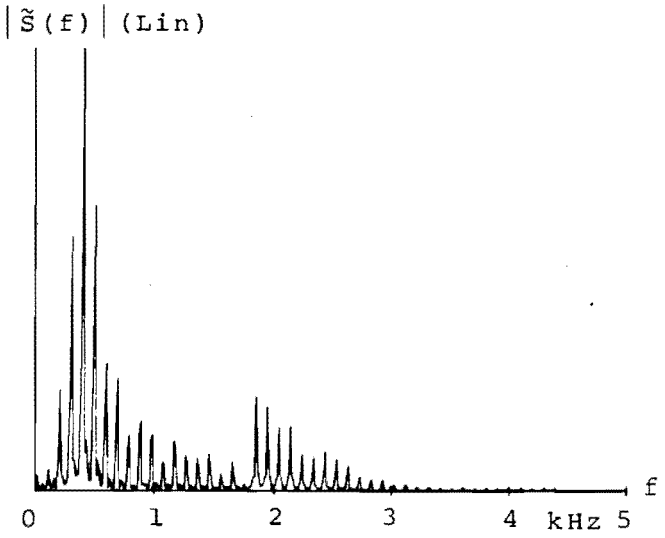


Fig 5.4.3

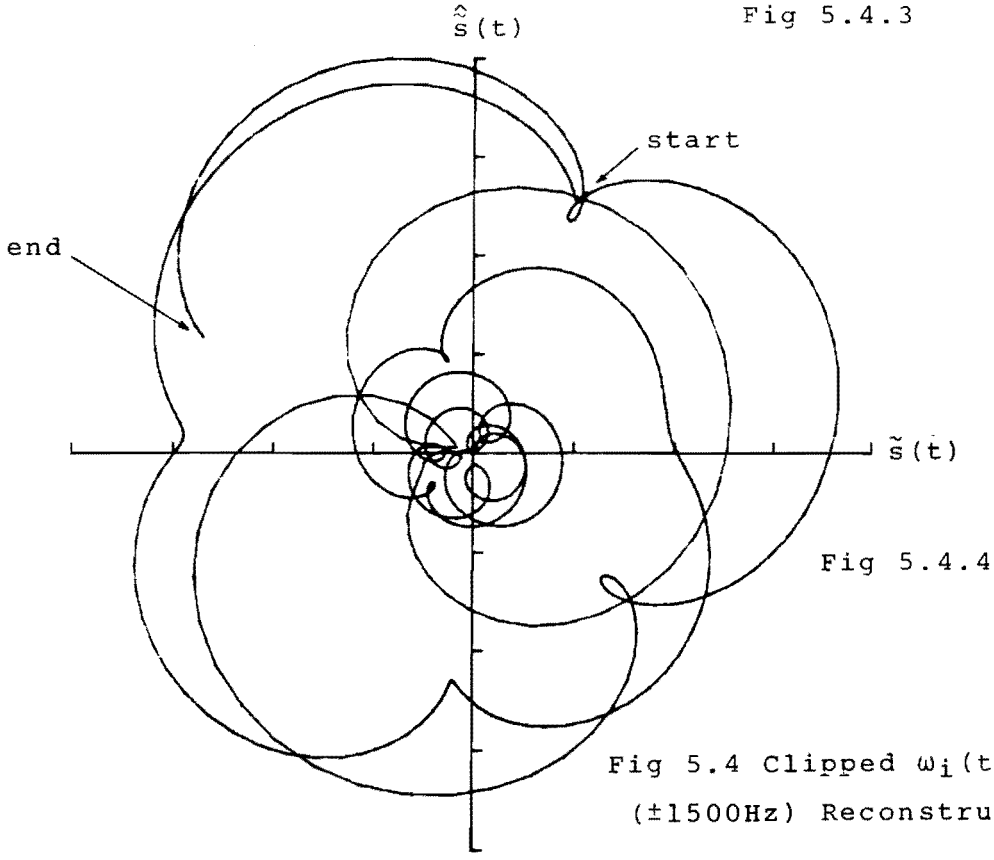


Fig 5.4.4

Fig 5.4 Clipped $\omega_i(t)$ (±1500Hz) Reconstruction

components. The most noticeable is distortion of the amplitude spectral envelope, but the associated frequency shift could also be significant in terms of speech intelligibility. The frequency shift is most easily thought of in terms of a change of average instantaneous frequency brought about by symmetrical clipping of the waveform $\omega_i(t)$ which is not symmetrical about 0 Hz.

The dynamic range of the computer based analysis system is sufficient for an apparently distortionless reconstruction of telephone bandwidth speech from instantaneous parameters stored at sampling rates of as low as 10,000 samples per second.

(5.3) FREQUENCY SHIFTING & DISTORTION ANALYSIS

The technique of frequency shifted reconstruction is particularly useful for investigating the spectral distortions produced by signal reconstruction from processed parameters.

If the real signal $s(t)$,

$$s(t) = a(t)\cos\{\int\omega_i(t).dt\} \qquad \ldots (5.3)$$

possesses the Fourier transform $S(f)$, then the amplitude and frequency modulating functions, $a(t)$ and $\omega_i(t)$, define the analytic signal $\Psi(t)$,

$$\Psi(t) = a(t)\exp\{j.\int\omega_i(t).dt\} \qquad \ldots (5.4)$$

which has the Fourier transform

$$\Psi(f) = \begin{cases} 2S(f), & f>0 \\ S(f), & f=0 \\ 0, & f<0 \end{cases} \qquad \ldots (5.5)$$

Examples of S(f) and $\Psi(f)$ are illustrated in figures (5.5.1) and (5.5.2). It can be seen that, through the analytic signal, a(t) and $\omega_i(t)$ describe the positive frequencies and bandwidth of s(t) exactly.

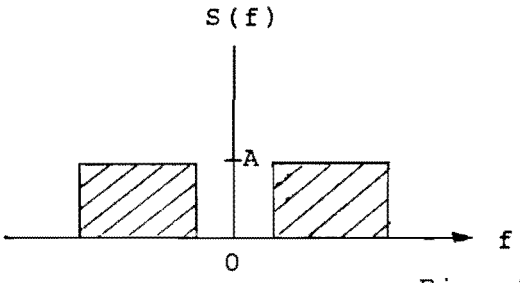


Fig. 5.5.1

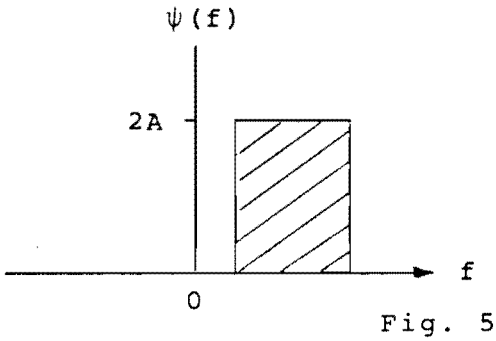


Fig. 5.5.2

Fig. 5.5 Real and Analytic Signal Spectra

Modifying one or both of a(t) or $\omega_i(t)$ leads to the new signal

$$\xi(t) = \tilde{a}(t)\exp\{j.\int\tilde{\omega}_i(t).dt\} \qquad \ldots (5.6)$$

which will usually exhibit wider bandwidth than $\Psi(t)$. (Ref.91) The spread bandwidth of one possible $\xi(t)$ is illustrated in figure (5.6.1) and the spectrum of the real signal resulting from the reconstruction

$$\tilde{s}(t) = \tilde{a}(t)\cos\{\int\tilde{\omega}_i(t).dt\} \quad \ldots (5.7)$$

is shown in figure (5.6.2).

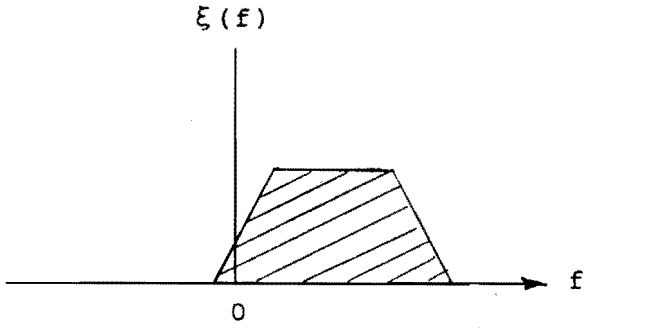


Fig. 5.6.1

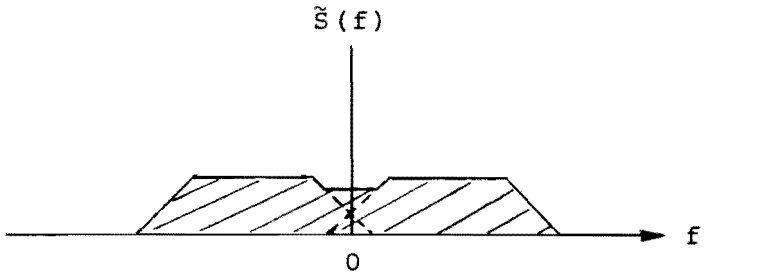


Fig. 5.6.2

Fig. 5.6 Modified Single Sided and Real Spectra

The spectrum of $\xi(t)$ has been visualised as spreading into negative Fourier frequencies and the resulting real reconstruction must therefore exhibit spectral overlap. The analytic signal corresponding to $\tilde{s}(t)$ is not $\xi(t)$ and instantaneous amplitude and frequency analysis of $\tilde{s}(t)$ would not result in the parameters $\tilde{a}(t)$ and $\tilde{\omega}_i(t)$.

The above example illustrates two stages in the distortion of real baseband signals reconstructed from modified instantaneous parameters. The first stage is distortion of the spectrum

of the associated analytic signal and this may cause $\tilde{s}(t)$ to exhibit greater bandwidth than s(t). The second stage is spectral overlap, or the folding of negative Fourier frequencies onto the low frequency end of the baseband reconstructions.

To remove the possibility of spectral overlap, the spectrum of the signal $\xi(t)$ may be shifted away from DC by adding a constant frequency value, ω_s, to its average instantaneous frequency.

$$\xi_s(t) = \tilde{a}(t)\exp\{\int(\tilde{\omega}_i(t)+\omega_s).dt\} \quad . . . (5.8)$$

This signal is illustrated in figure (5.7.1) along with its associated real reconstruction

$$\tilde{s}_s(t) = \tilde{a}(t)\cos\{\int(\tilde{\omega}_i(t)+\omega_s).dt\} \quad . . . (5.9)$$

in figure (5.7.2)

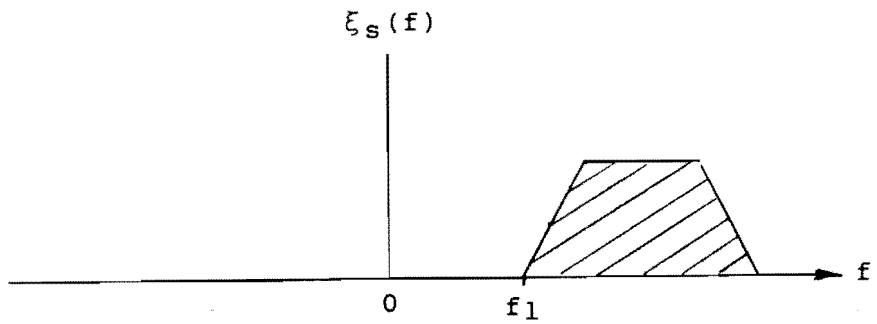


Fig. 5.7.1

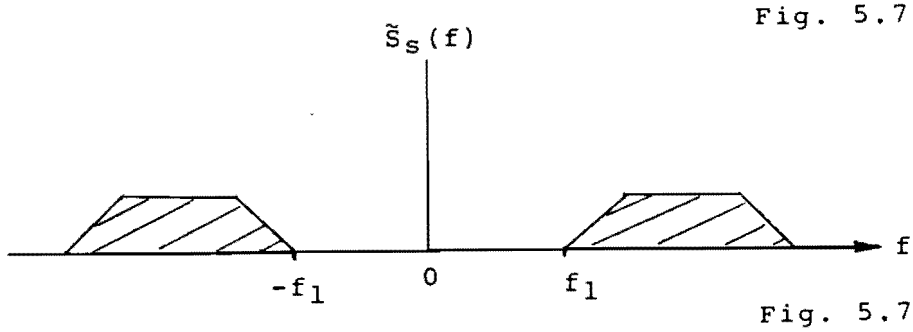


Fig. 5.7.2

Fig. 5.7 Frequency Shifted Reconstructions

Examination of the spectrum of $\tilde{s}_s(t)$ reveals any spectral spreading which has occurred and any possibility of spectral overlap at baseband. Spectral overlap will occur if $f_1 < \omega_s/2\pi$.

A frequency shifted reconstruction of the bandpass vowel /ε/ (figure (5.2)) according to

$$\tilde{s}(t) = a(t)\cos\{\int(\omega_i(t)+2\pi x12{,}000).dt\} \quad . . . \quad (5.10)$$

is illustrated in figure (5.8). Predictably, the frequency shifted reconstruction exhibits no spectral components below 12,000 Hz.

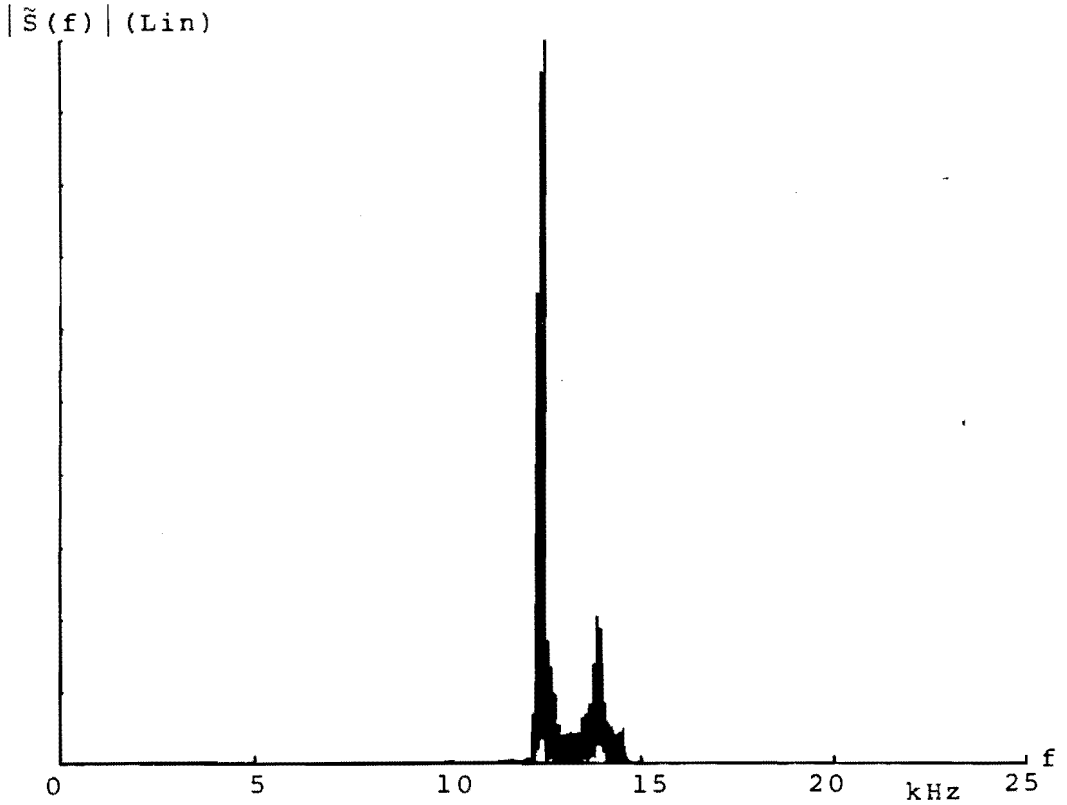


Fig 5.8 Frequency Shifted Reconstruction of /ε/

The analytic signal $\gamma(t)$ which possesses the Fourier transform

$$\gamma(f) = \begin{cases} 0, f>0 \\ S(f), f=0 \\ 2S(f), f<0 \end{cases} \quad . . . \quad (5.11)$$

is defined by the instantaneous parameters $a(t)$ and $-\omega_i(t)$. Reconstruction of a real signal at baseband

$$\tilde{s}(t) = a(t)\cos\{-\int\omega_i(t).dt\} \quad . . . \quad (5.12)$$

is identical to reconstruction with a(t) and $+\omega_i(t)$ because of the real signals spectral symmetry about DC. A frequency shifted real reconstruction, however, may generate a spectrally inverted version of the original signal, if ω_s is greater than the signal bandwidth. This is also illustrated for the bandpass vowel /ɛ/ in figure (5.9), using the reconstruction

$$\tilde{s}(t) = a(t)\cos\{\int((-\omega_i(t))+2\pi x12{,}000).dt\} \quad . \; . \; . \; (5.13)$$

In this case, the reconstruction exhibits no spectral components above 12,000 Hz.

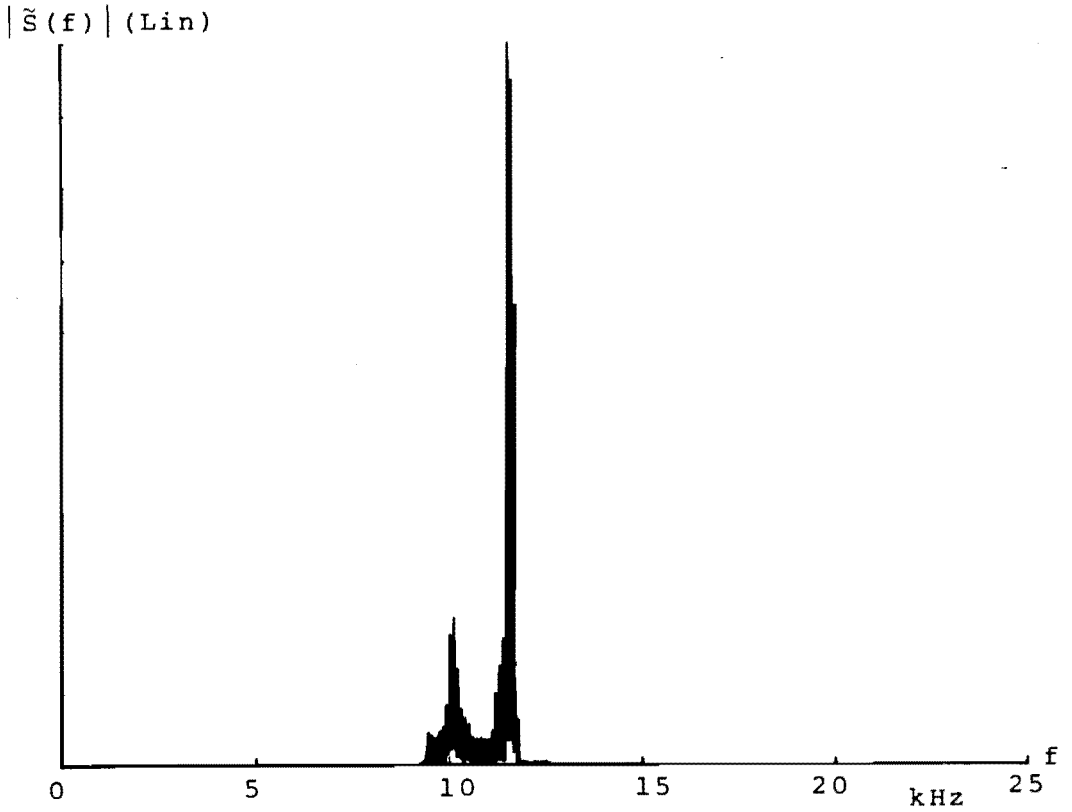


Fig 5.9 Frequency Inverted and Shifted Reconstruction of /ɛ/

Voelker's equations (2.61) and (2.62) show that the analytic signal zeros of $\gamma(z)$ are the complex conjugates of those of $\Psi(z)$ (equation (5.4)). Spectral inversion, therefore corresponds to conversion of all analytic signal LHP zeros to UHP and vice-versa. The spectrally inverted signal $\gamma(z)$ is the complex conjugate of $\Psi(z)$.

(5.4) CONSTANT AMPLITUDE RECONSTRUCTIONS

Except for multiplicative constants, a full definition of an analytic signal is afforded by the locations of its zeros in real and imaginary time (Section (4.3.1)). The fact that both the magnitude and argument of these zero locations are required to generate the instantaneous frequency waveform, equation (2.62), suggests that a constant amplitude reconstruction

$$\tilde{s}(t) = \cos\{\int \omega_i(t).dt\} \qquad \ldots (5.14)$$

may retain many of the features of the original real signal, s(t).

Constant amplitude speech signals, equivalent to that described by equation (5.14), have been generated by SSB modulation techniques (Ref. 94) and found to be "highly intelligible". Signal quality is reduced, however, by disagreeable noise during inter-phoneme and inter-word silences.

The types of signal distortion introduced by constant amplitude reconstruction may be illustrated using the bandpass vowel /ɛ/ and frequency shifting techniques. Figure (5.10) shows the amplitude spectrum, real waveform and instantaneous parameters of the baseband vowel. Performing a constant amplitude frequency shifted reconstruction

$$\tilde{s}(t) = \cos\{(\int \omega_i(t) + 2\pi \times 12{,}000).dt\} \qquad \ldots (5.15)$$

results in the amplitude spectrum figure (5.11). This may be compared with the baseband spectrum, frequency shifted by the same factor, figure (5.12).

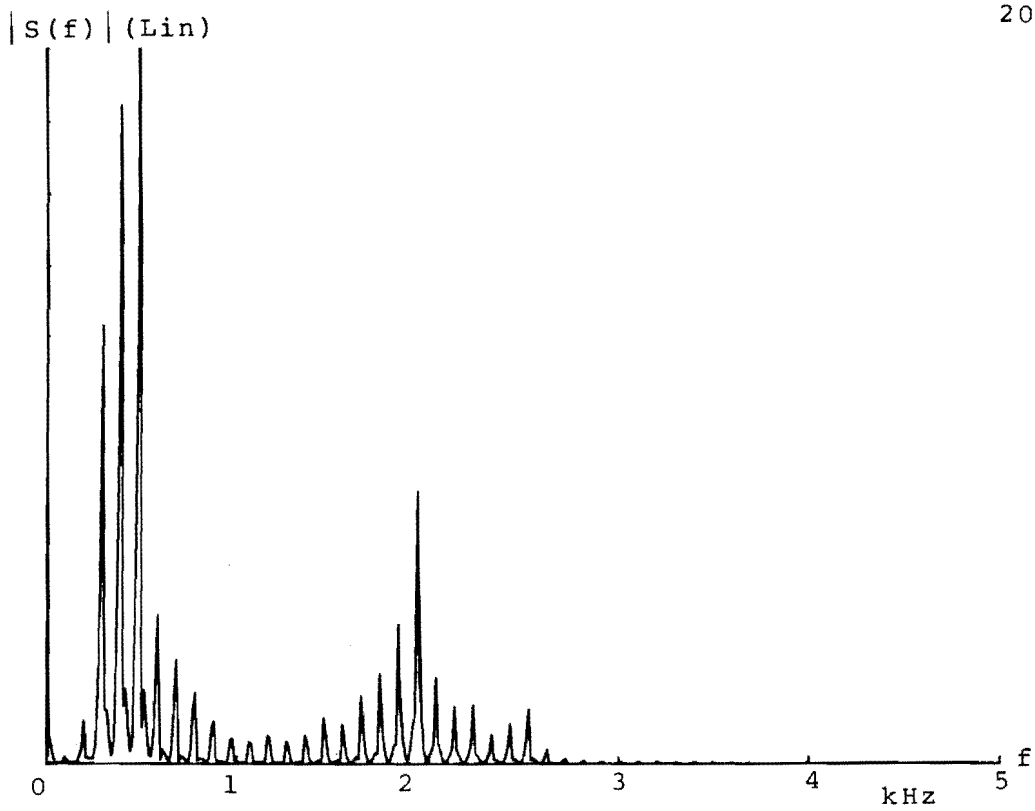


Fig 5.10.1

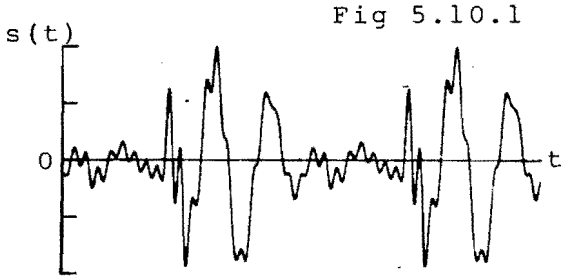


Fig 5.10.2

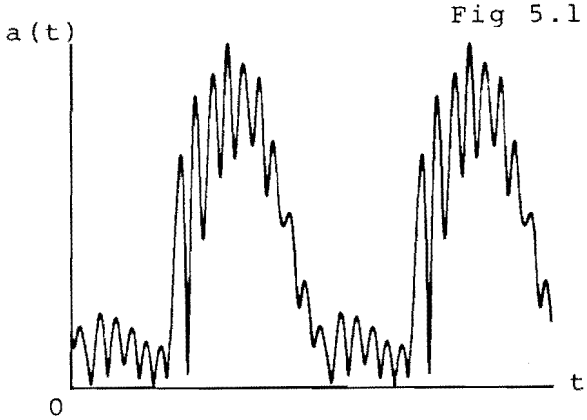


Fig 5.10.4

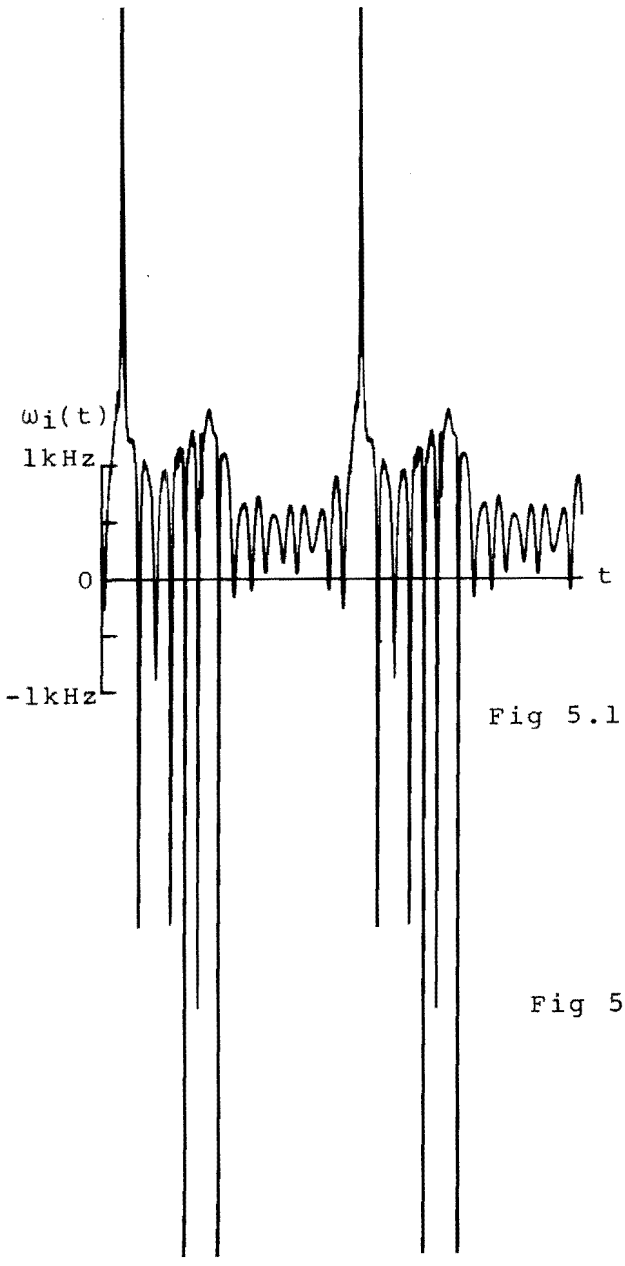


Fig 5.10.3

Fig 5.10 Reference Vowel /ɛ/

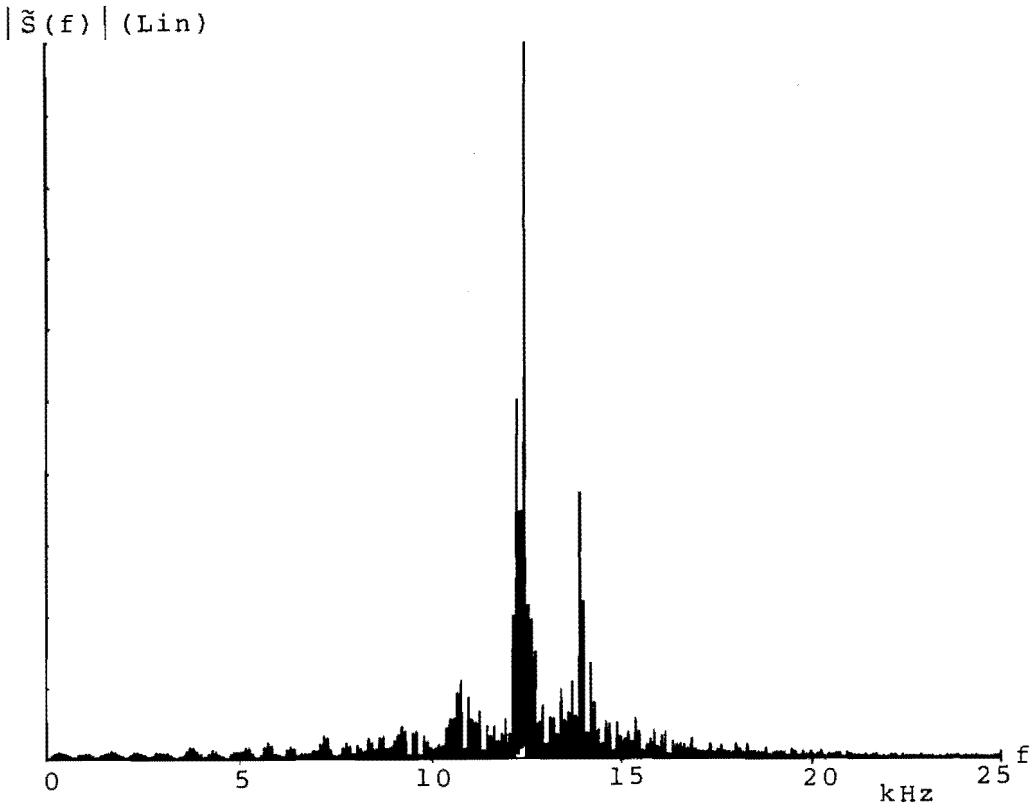


Fig 5.11 Constant Amplitude Frequency Shifted Reconstruction of /ɛ/

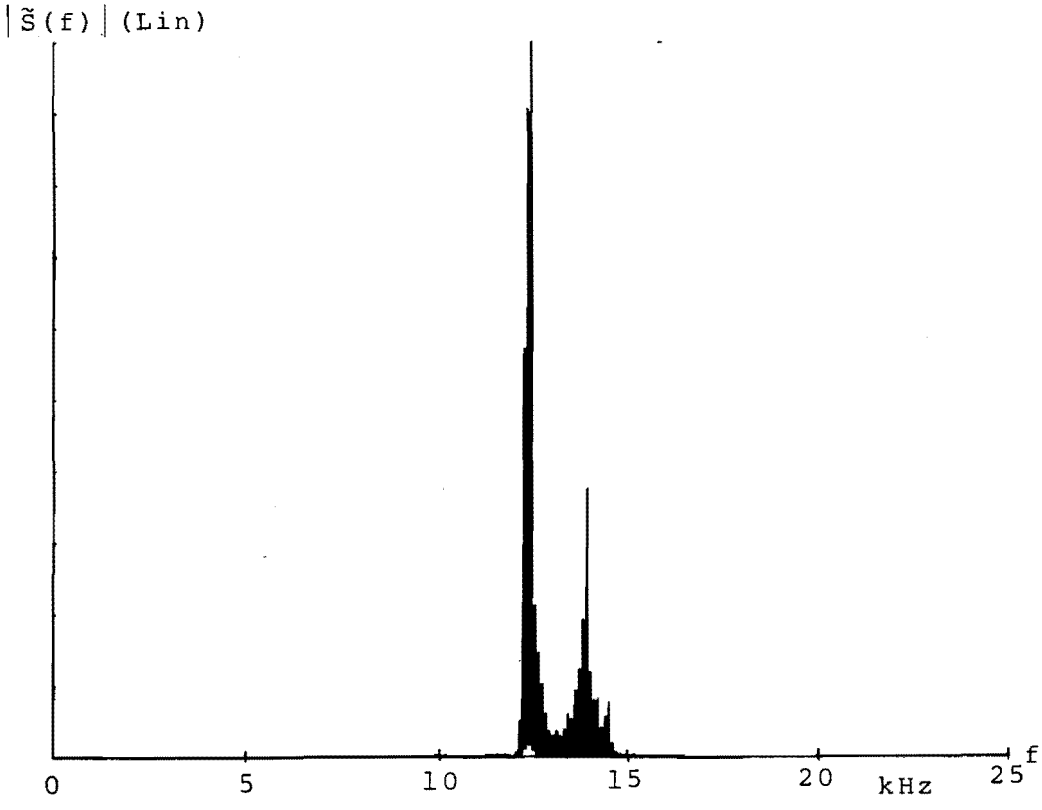


Fig 5.12 Frequency Shifted Reconstruction of /ɛ/

Figure (5.11) shows thatthe amplitude spectrum line structure and first and second formant positions are preserved. The signal bandwidth is greatly increased, however, as the bandwidth preserving relationship between a(t) and $\omega_i(t)$ has been destroyed. As there is no corresponding instantaneous amplitude dip, large instantaneous frequency excursions now cause high modulation index frequency modulation.

In accordance with frequency modulation theory, removal of the amplitude modulating function has resulted in sidebands being more evenly distributed around the carrier frequency $\overline{\omega_i(t)}$. Instantaneous frequency is generally not symmetrical around $\overline{\omega_i(t)}$, however, and the "image" second formant created at a frequency of approximately $(\overline{\omega_i(t)}-\omega_{f2})+\omega_s$ is of reduced amplitude (ω_{fn} is the centre frequency of the n^{th} formant and ω_s is the reconstruction frequency shift).

Frequency shifting the constant amplitude reconstruction amplitude spectrum back to baseband would cause spectral overlap and the "image" second formant would fold from negative Fourier frequencies to approximately $(\omega_{f2}-\overline{\omega_i(t)})-\overline{\omega_i(t)}$. This frequency is normally between ω_{f1} and ω_{f2}.

Figure (5.13) is the amplitude spectrum and waveform of the full bandwidth vowel /ɛ/. Constant amplitude reconstruction at baseband results in the spectrum, figure (5.14.1), which exhibits the expected increased bandwidth and "image" components between ω_{f1} and ω_{f2}. However, as the basic line and formant structure of the amplitude spectrum is maintained, the vowel remains intelligible. The constant amplitude time waveform is illustrated in figure (5.14.2).

Previous investigations have shown that the amplitude spectrum of a bandlimited unvoiced phoneme may be considered to be almost symmetrical about its average instantaneous frequency. (Section (4.3.2.4)).

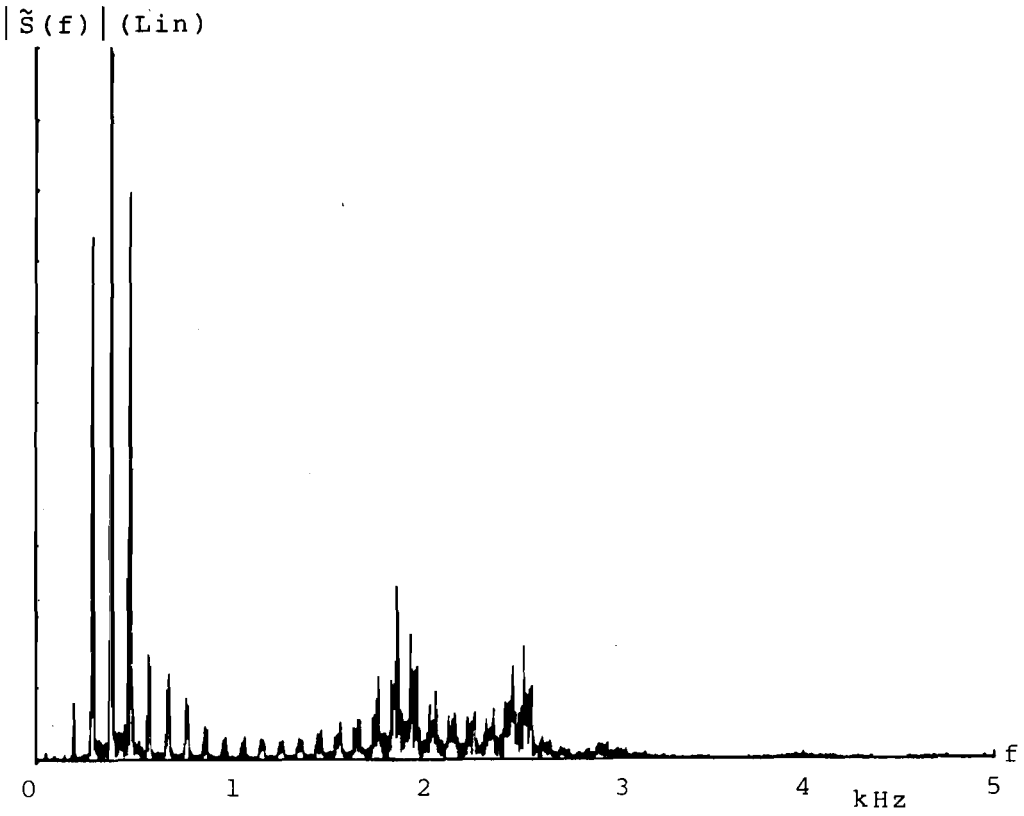


Fig 5.13.1

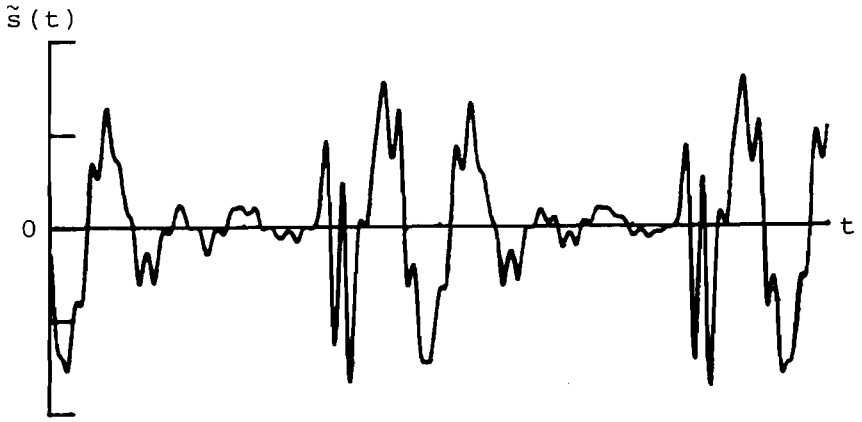


Fig 5.13.2

Fig 5.13 Undistorted Reconstruction of /ε/

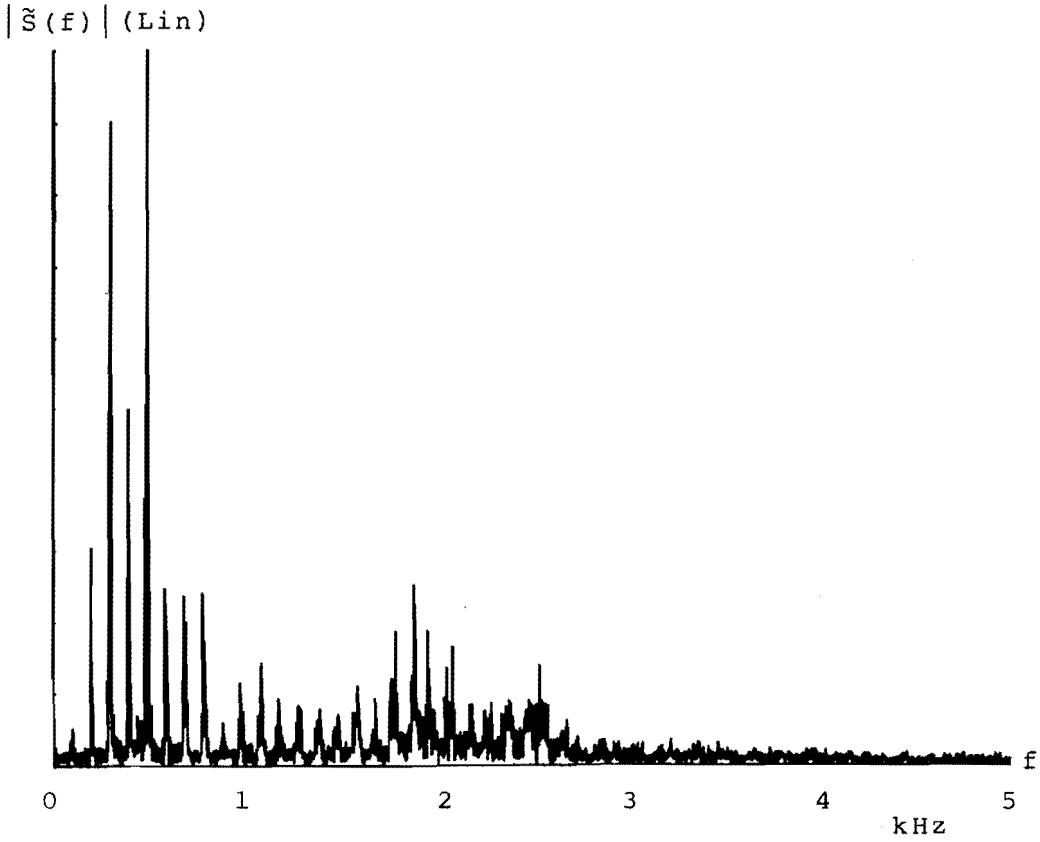


Fig 5.14.1

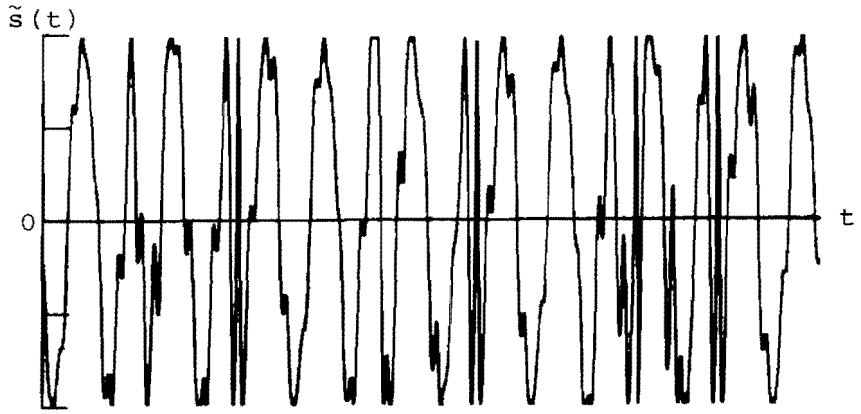


Fig 5.14.2

Fig 5.14 Constant Amplitude Reconstruction of /ɛ/

If this is the case, constant amplitude reconstruction of unvoiced fricatives will cause a bandwidth increase but few "image" resonances.

(5.4.1) CLIPPED SPEECH

The intelligibility of constant amplitude speech and the surprisingly high intelligibility of hard limited or clipped speech, $S_c(t)$ are closely related. Clipped speech may be represented by the Fourier series

$$S_c(t) = \sum_n (1/n)\cos\{n \times \int \omega_i(t).dt\}, n=1, 3, 5.... \qquad (5.16)$$

which is essentially constant amplitude speech plus high frequency distortion. The frequency multiplied signals which make up the high frequency components of $S_c(t)$ are considered in more detail, in Section (5.5), but may be approximated by frequency shifted versions of $\cos\{\int \omega_i(t).dt\}$ with average instantaneous frequencies $n \times \overline{\omega_i(t)}$. The amplitude spectrum of a clipped vowel would therefore be expected to display the vowels formant and line structure, plus an "image" second formant at frequency $(\omega_{f2} - \overline{\omega_i(t)}) - \overline{\omega_i(t)}$, and reduced amplitude (ratio 1/n) first formant images at approximately $n \times \overline{\omega_i(t)}$ with associated upper and lower second formants at $n \times \overline{\omega_i(t)} \pm (\omega_{f2} - \overline{\omega_i(t)})$. Only a few of the frequency multiplied formants and images fall within the baseband.

The effects of clipping are illustrated for a bandpass vowel whose amplitude spectrum, waveform, instantaneous parameters and vector plot are shown in figure (5.15). the amplitude spectrum of the waveform after clipping, figure (5.16.1) displays severe distortion of the first and second formants and reveals the presence of a strong image component between ω_{f1} and ω_{f2}. The increased bandwidth is consistent with the sharp transitions of the clipped time waveform figure (5.16.2)

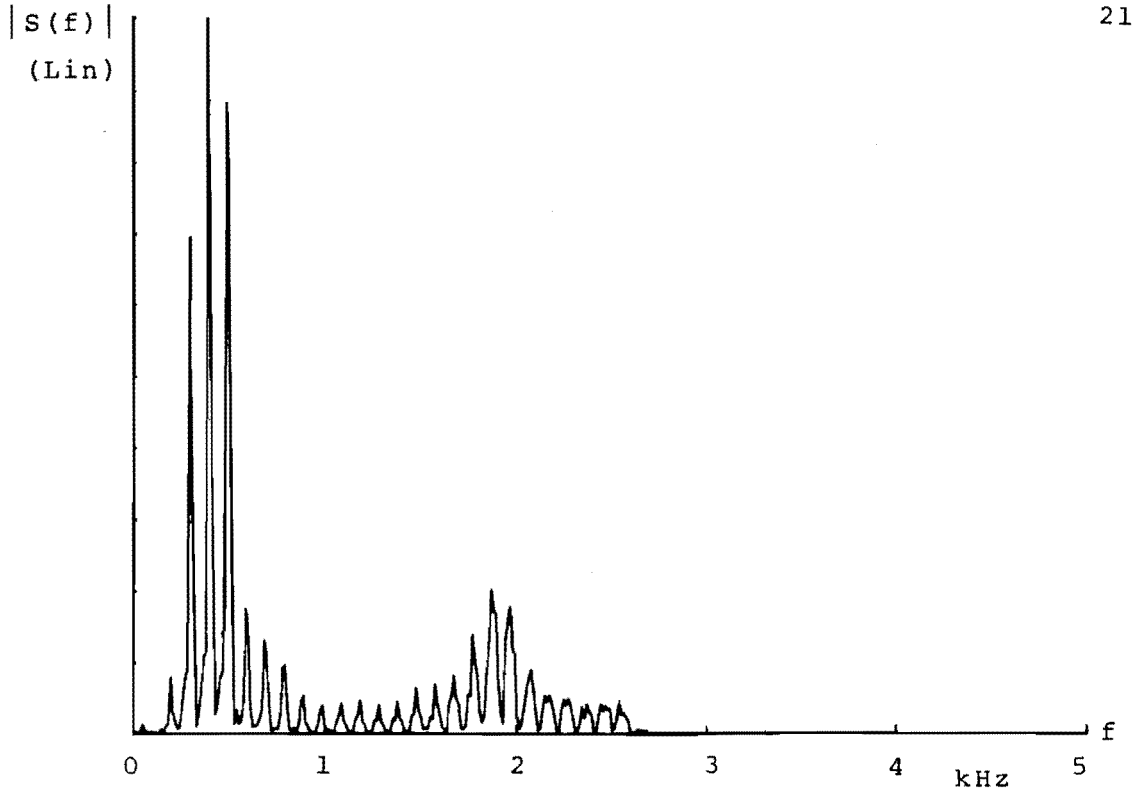


Fig 5.15.1

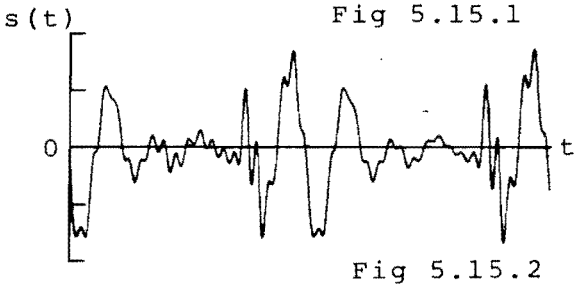


Fig 5.15.2

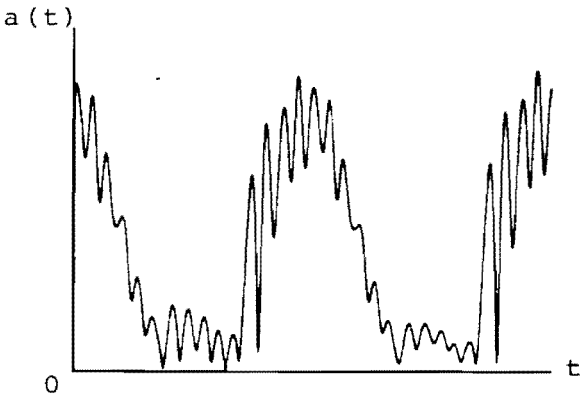


Fig 5.15.4

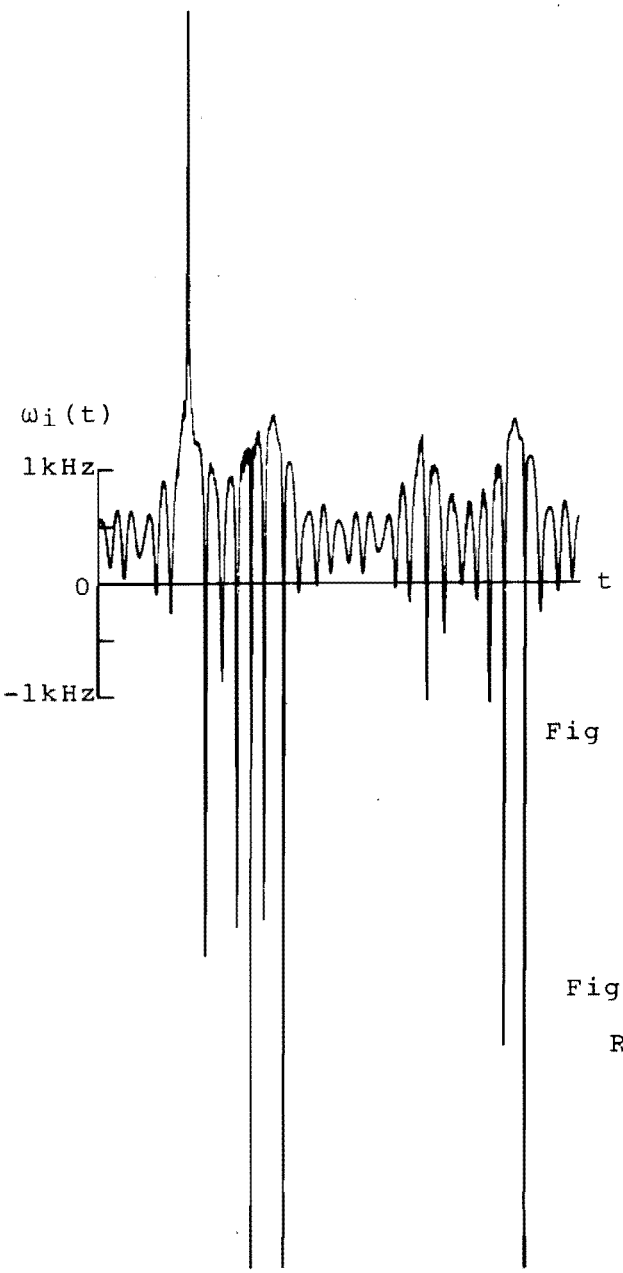


Fig 5.15.3

Figs 5.15.1-5.15.4

Reference Waveforms for Clipping and Analysis

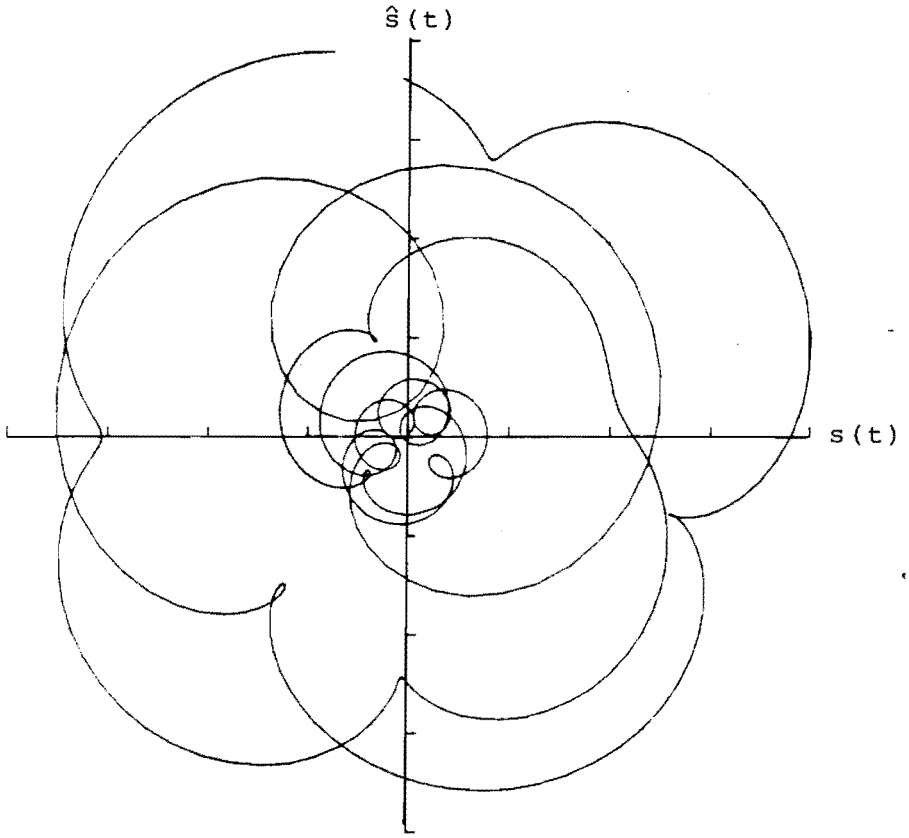


Fig 5.15.5 Reference Waveforms for Clipping and Analysis

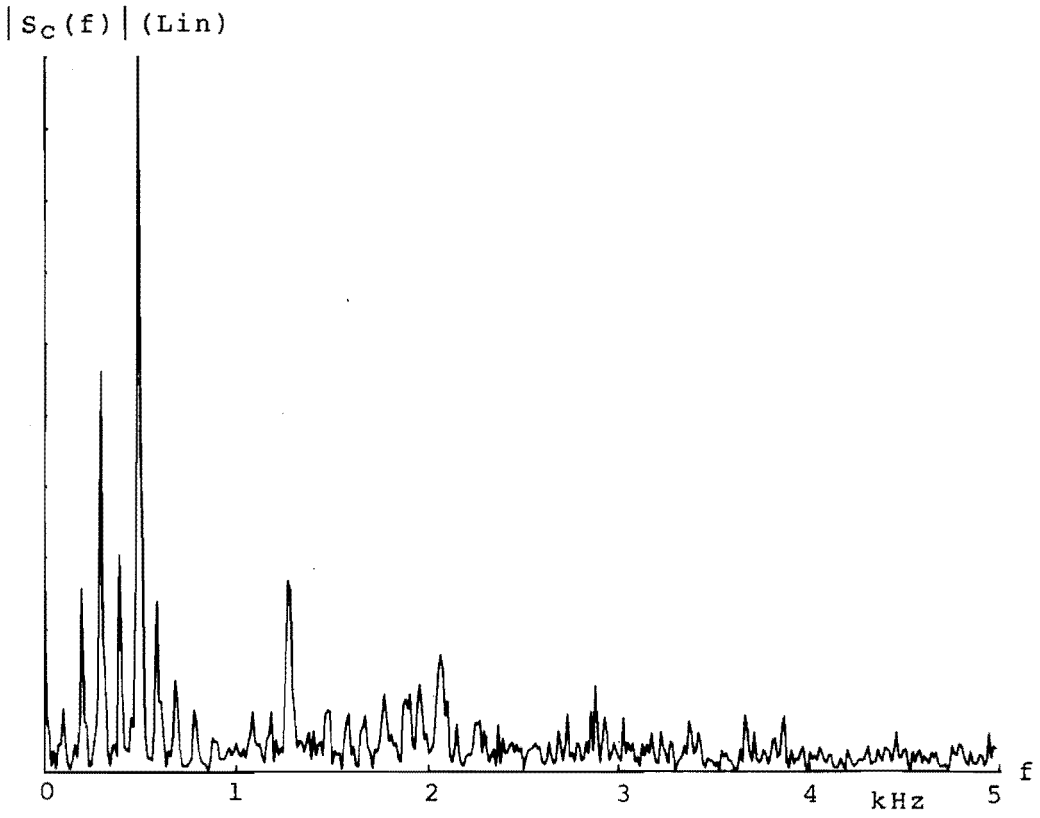


Fig 5.16.1

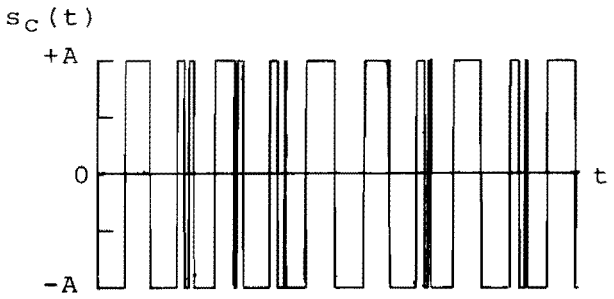


Fig 5.16.2

Figs 5.16.1-5.16.2 Clipped Reference Vowel; Amplitude Spectrum and Time Waveform

The instantaneous amplitude of the clipped signal is constant at +A and its instantaneous frequency is zero with a positive impulse (area π radians) at each zero crossing of $S_c(t)$. The average instantaneous frequency is therefore $\pi.n_{zc}$ radians per second where n_{zc} is the number of zero crossings per cycle. The average, $\pi.n_{zc}$, is generally not equal to the average instantaneous frequency of the real signal as some of the zero crossings of s(t) may be due to negative instantaneous frequency excursions and their associated vector inner loops.

In order to perform an instantaneous parameter analysis of the clipped waveform, it is necessary to lowpass filter to 3,500 Hz. This causes the waveform to "ring" (Gibb's phenomenon), and the resulting low pass signal, figure (5.16.3), has the instantaneous functions figures (5.16.4) and (5.16.5).

By reducing the speed at which the clipped waveform may cross the time axis, lowpass filtering has introduced significant instantaneous amplitude dips and has "spread" the instantaneous frequency impulses. The instantaneous amplitude and frequency dips associated with waveform "ringing" correspond to the few LHP analytic signal zeros which remain after the lowpass operation.

Comparison of this instantaneous frequency curve with that of the original vowel, figure (5.15.3), confirms that several LHP analytic signal zeros have been converted to UHP by the clipping operation.

Figure (5.16.6) is a version of the clipped signal generated by reconstruction from the instantantous parameters of the unclipped signal according to

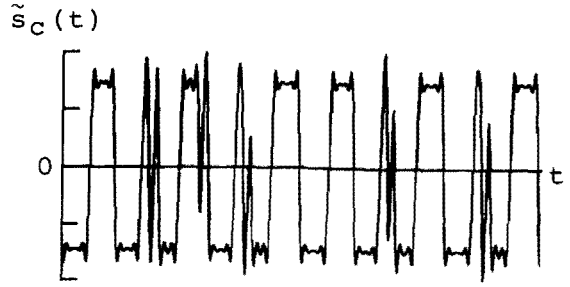


Fig 5.16.3

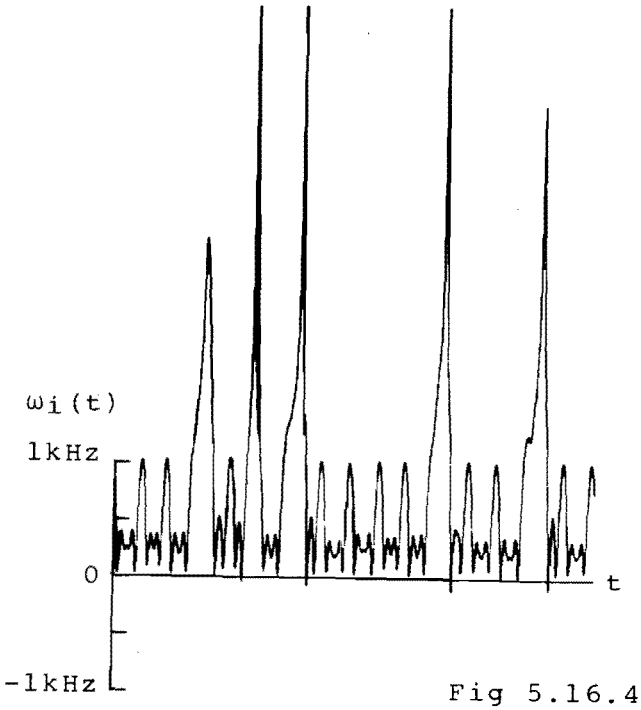


Fig 5.16.4

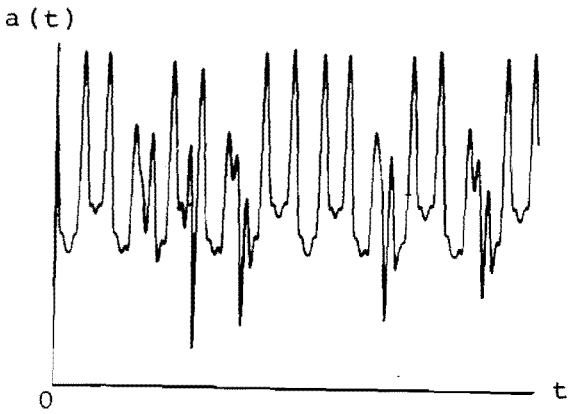


Fig 5.16.5

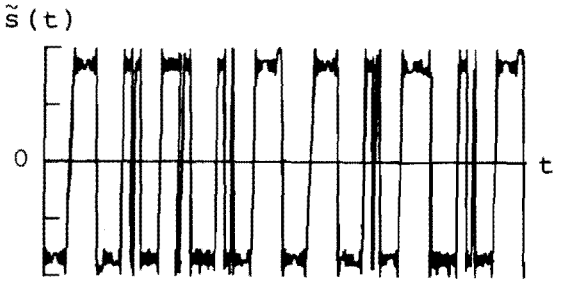


Fig 5.16.6

Figs 5.16.3-5.16.6 Instantaneous Parameter Analysis of Lowpass $s_c(t)$

$$\tilde{s}(t) = \sum (1/n)\cos\{nx\int\omega_i(t).dt\}, n=1,3,...11. \qquad (5.17)$$

This resembles a lowpass version of $S_C(t)$ and confirms the relationship between constant amplitude and clipped speech.

The average instantaneous frequency of an unvoiced phoneme is normally high (>1,000 Hz) and the high frequency distortion components caused by clipping ($(1/n)\cos\{nx\int\omega_i(t).dt\}$) fall out of the baseband. For this reason, clipping of unvoiced phonemes results in less noticeable distortion than clipping of voiced phonemes.

Investigation into the properties of clipped speech have revealed that differentiation of the voice signal prior to clipping greatly improves the intelligibility of $S_C(t)$ (Ref. 83). This phenomenon can be successfully explained by the fact that differentiation converts some complex conjugate pairs of real signal zeros, into pairs of zero crossings. As the complex and real zeros may be considered to be "informational attributes" of the real signal, more "information" is preserved by clipping after differentiation than by straightforward clipping.

Although highly intelligible, differentiated then clipped speech still exhibits some distortion and this may be illustrated by a differentiated version of the bandpass vowel used to describe constant amplitude speech. Figures (5.17.1) and (5.17.2) are the amplitude spectrum and time waveform of the differentiated vowel. Predictably, differentiation has increased the magnitude of the second formant and the waveform exhibits many new zero crossings. The instantaneous parameters, figures (5.17.3) and (5.17.4) are significantly different from those of the original vowel and the vector plot, figure (5.17.5), reveals that average instantaneous frequency now corresponds to that of the 15th harmonic. This indicates that the differentiation process has converted 10 LHP analytic signal complex zeros

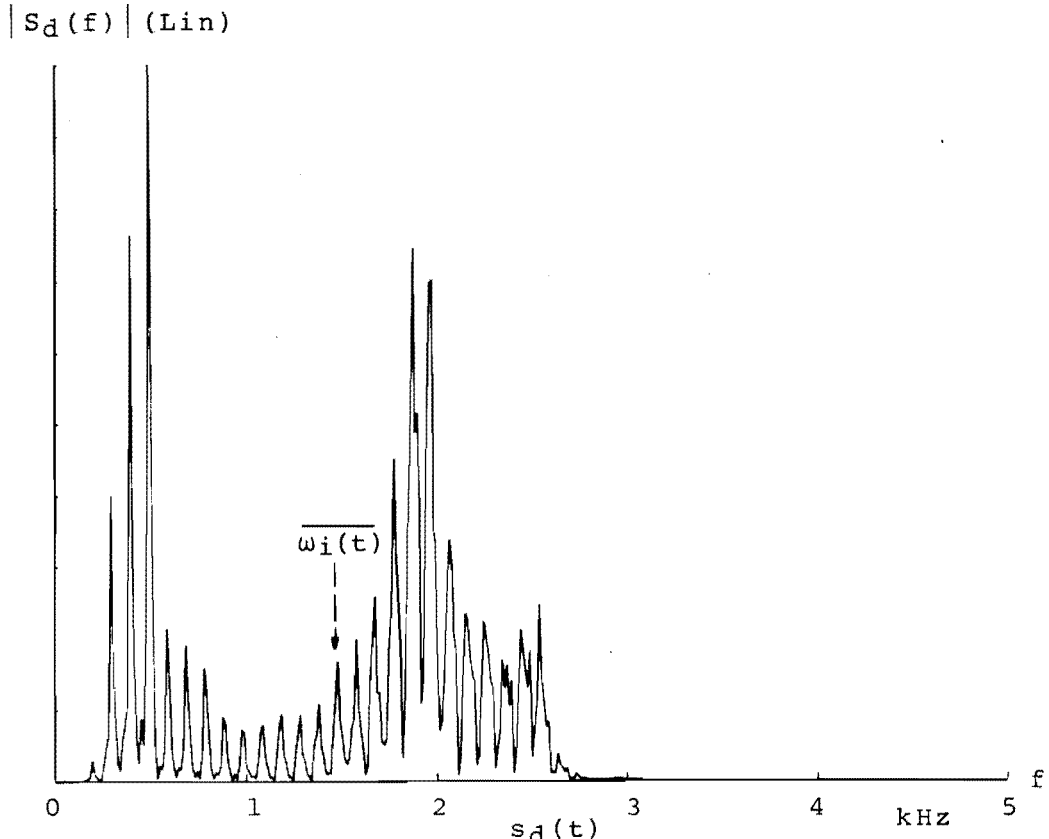


Fig 5.17.1

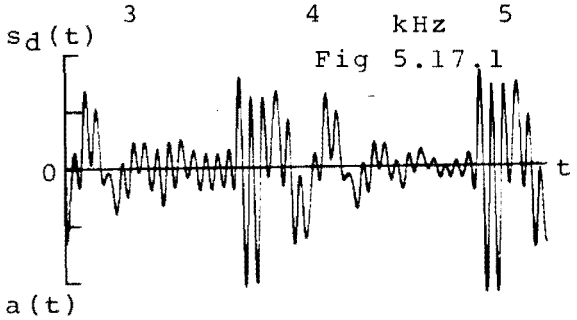


Fig 5.17.2

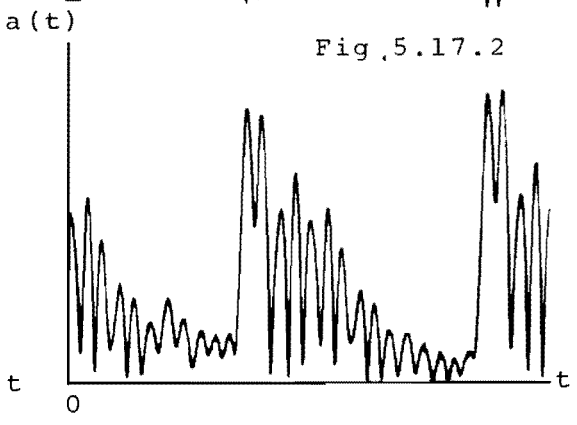


Fig 5.17.4

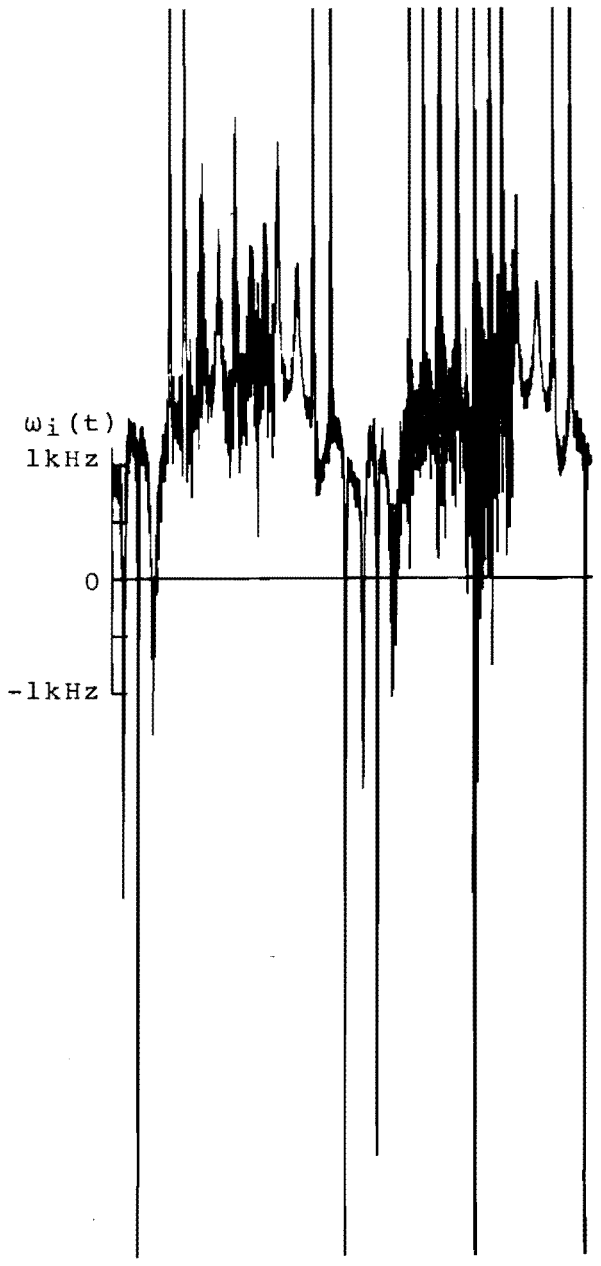


Fig 5.17.3

Figs 5.17.1-5.17.4
Analysis of Differentiated Vowel

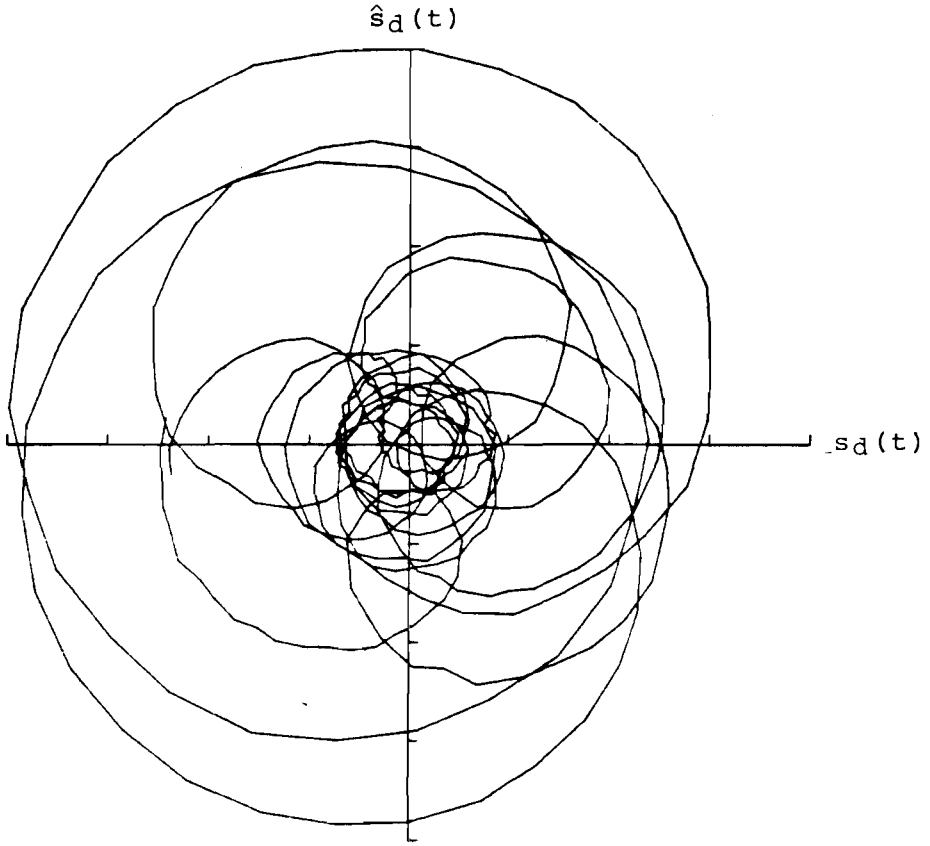


Fig 5.17.5 Analysis of Differentiated Vowel

per cycle to UHP.

The clipped version of the differentiated signal can be represented

$$S_c(t) = \sum_n (1/n)\cos\{nx\int\omega_i(t).dt\}, n=1,3,5... \qquad (5.18)$$

The condition n=1 corresponds to the constant amplitude reconstruction

$$\tilde{s}(t) = \cos\{\int\omega_i(t).dt\} \qquad ... \quad (5.19)$$

which, for the case of the above vowel, will exhibit distortion consisting of two reduced amplitude "image" formants. The image first formant is at frequency $\overline{\omega_i(t)}+(\overline{\omega_i(t)}-\omega_{f1})$ and the image second at $\overline{\omega_i(t)}-(\omega_{f2}-\overline{\omega_i(t)})$. Once again, these images are due to symmetricalisation of the F.M. spectrum around the carrier, $\overline{\omega_i(t)}$.

The high frequency distortion components $(1/n)\cos\{nx\int\omega_i(t).dt\}$ have average instantaneous frequency $nx\overline{\omega_i(t)}$. As the differentiation process has raised $\overline{\omega_i(t)}$ to approximately 2πx1,500 radians per second, all of these distortion components are essentially thrown out of the baseband and may be removed by lowpass filtering to 3,400 Hz. The resulting signal can be integrated to produce a vowel distorted only by low amplitude image formants.

Figures (5.18.1) and (5.18.2) are the amplitude spectrum and time waveform of the differentiated, then clipped vowel. The average instantaneous frequency (carrier) and image first and second formants are indicated on the amplitude spectrum. Apart from the presence of some higher frequency noise and "image" formants, figure (5.18.1) closely resembles the spectrum of the unclipped signal, figure (5.17.1).

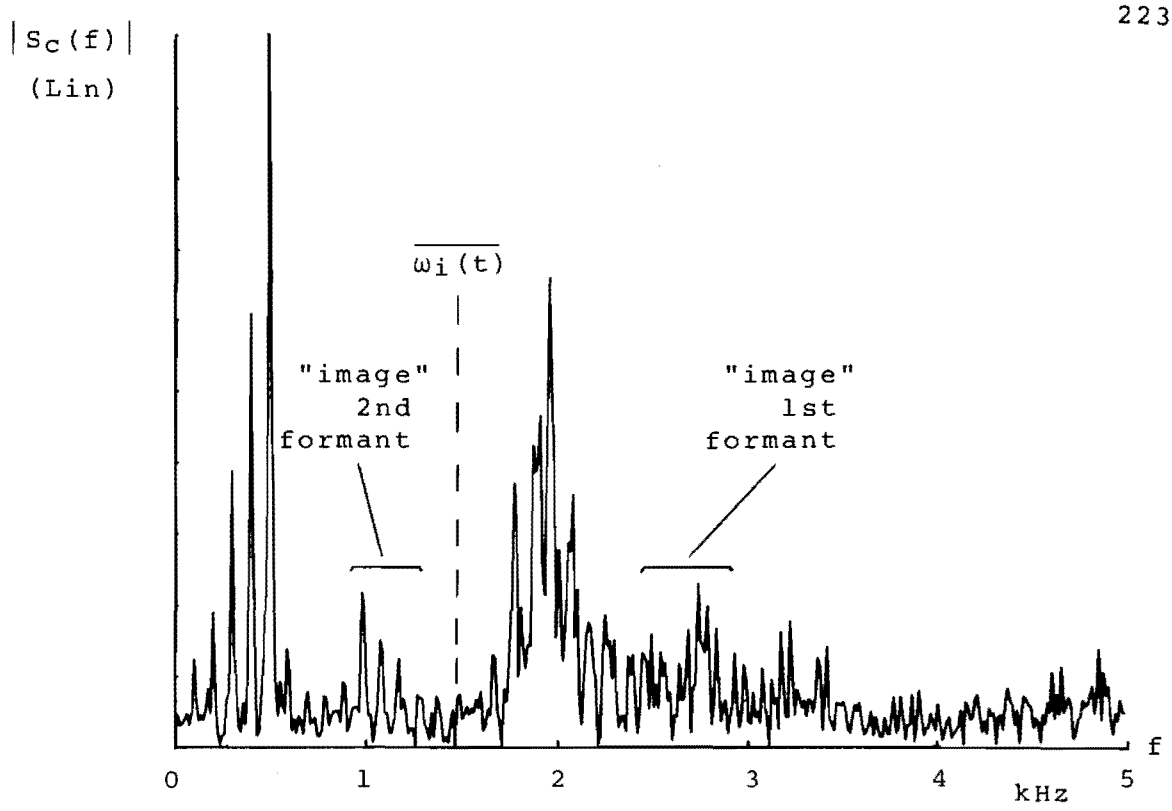


Fig 5.18.1

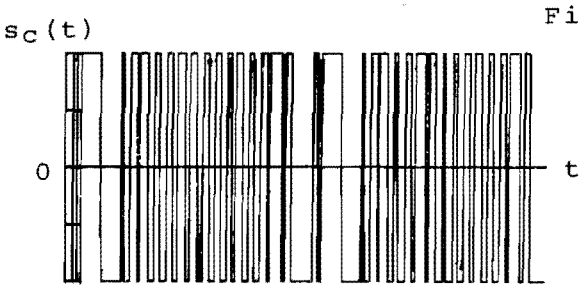


Fig 5.18.2

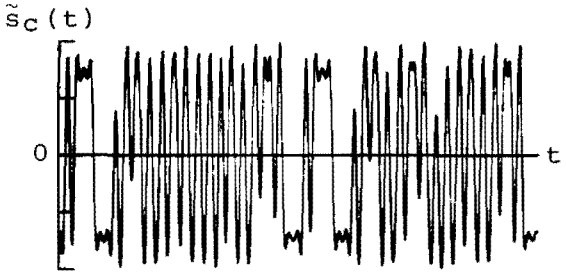


Fig 5.18.3

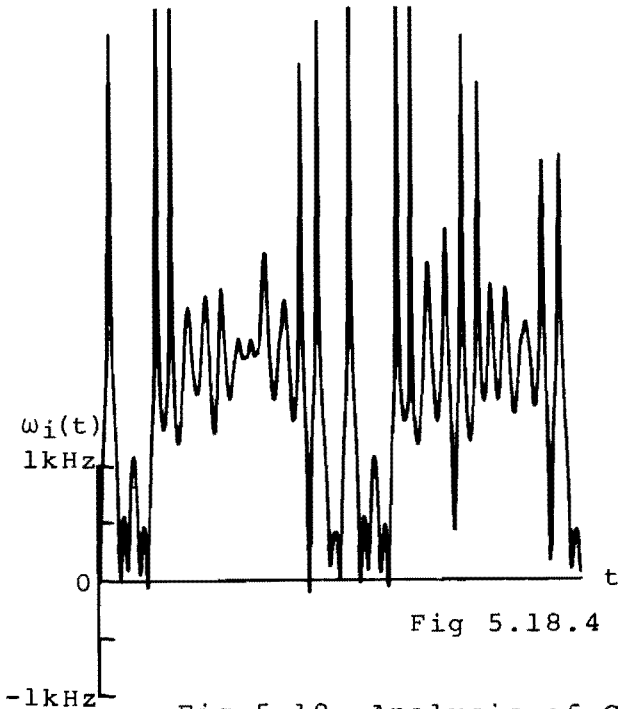


Fig 5.18.4

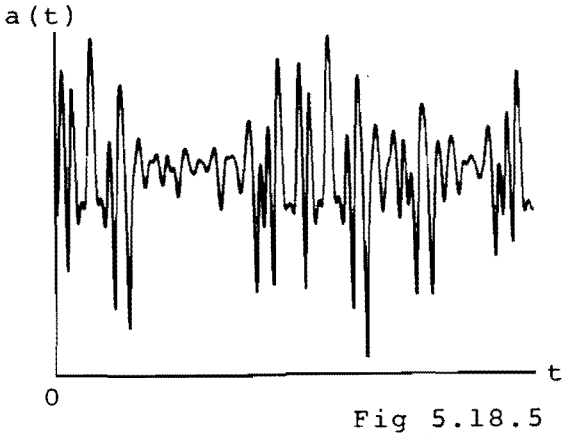


Fig 5.18.5

Fig 5.18 Analysis of Clipped Differentiated Vowel

Instantaneous parameter analysis of the clipped waveform low pass filtered to 3,400 Hz, figure (5.18.3), results in the instantaneous waveforms, figures (5.18.4) and (5.18.5). The instantaneous frequency curve, figure (5.18.4), resembles that of the differentiated, but unclipped signal, figure (5.17.3), as the differentiated signal possesses a high proportion of UHP analytic signal zeros. Filtering a clipped signal to the bandwidth of the original generates instantaneous amplitude dips and instantaneous frequency rises consistent with some UHP analytic signal zeros of the original signal. (Section (5.8)). The LHP analytic signal zero structure, however, is not re-generated by simple bandlimitation and LHP zeros indicated by figures (5.18.4) and (5.18.5) are due to amplitude ripple (Gibb's phenomenon).

All of the analytic signal zeros of a speech waveform may be made UHP by the addition of a high amplitude sinusoid at frequency $\omega_{sin} > 2\pi \times 3{,}500$ radians per second (for telephone bandwidth voice). This "bias" signal must be of sufficiently high amplitude that $\overline{\omega_i(t)} = \omega_{sin}$. Constant amplitude reconstruction of the "voice plus bias" signal leads to the creation of "image" formants above the carrier frequency, ω_{sin}, and there is no distortion of the lowpass speech amplitude spectrum.

Similarly, clipping of the "voice plus bias" signal creates the "image" formants above ω_{sin} and the lowest frequency distortion component, $(1/n)\cos\{n \times \int \omega_i(t).dt\}$, is centred around a carrier of $3 \times \omega_{sin}$. Lowpass filtering of the clipped signal to 3,400 Hz restores the undistorted speech signal.

The "voice plus bias" signal, where $\overline{\omega_i(t)} = \omega_{sin}$, is the basis of recent research into a signals sinewave crossings (Ref. 117) and signals designed for recovery after clipping (Ref. 118).

(5.5) FREQUENCY DIVISION AND MULTIPLICATION

The prospect of bandwidth reduction through division of the instantaneous frequency waveforms of sub-bands of speech has been stimulation for the design of several types of vocoder (Ref. 55). The basic premise is that frequency division of a particular sub-band (which holds at most one formant of a voiced sound) by a factor of n, allows transmission of that sub-band over $(1/n)^{th}$ of its original bandwidth. Unfortunately, research into the shortcomings of frequency division vocoders has shown that frequency division causes a frequency shift, but no identifiable bandwidth compression (Ref. 63).

Studies of the relationships between the amplitude spectra and instantaneous frequency waveforms for several phonemes and knowledge of the required relationships between $a(t)$ and $\omega_i(t)$ for distortionless reconstruction, allow the definition of distortions caused by frequency division or multiplication.

The reconstruction

$$\tilde{s}_n(t) = a(t)\cos\{(\int\omega_i(t).dt)/n\} \quad \ldots (5.20)$$

where $\tilde{s}_n(t)$ is a frequency divided sub-band of speech, has average instantaneous frequency $\overline{\omega_i(t)}/n$ and therefore displays a frequency shift of $((1-n)/n)\times\overline{\omega_i(t)}$. The magnitudes of instantaneous frequency excursions have also been divided by n, thus ensuring that the constant amplitude reconstruction

$$\tilde{g}_n(t) = \cos\{(\int\omega_i(t).dt)/n\} \quad \ldots (5.21)$$

exhibits less bandwidth than

$$\tilde{g}(t) = \cos\{\int\omega_i(t).dt\} \quad \ldots (5.22)$$

If the frequency division factor, n, is sufficiently large, instantaneous frequency fluctuations may be reduced to the extent that the amplitude spectrum of $\tilde{s}_n(t)$ resembles simple amplitude modulation of the carrier at $\overline{\omega_i(t)}/n$ by a(t). Assuming that the original sub-band contained a single vowel formant, such symmetricalisation of the sub-band signal spectrum may be a negligible distortion effect.

The rate at which instantaneous amplitude and frequency fluctuations occur is determined by the rate of occurrence of analytic signal zeros. Division of the instantaneous frequency waveform does not alter the rate of fluctuation occurrence and cannot affect any periodic properties of the instantaneous waveform. For this reason, frequency division does not alter the spectral line spacing of a periodic signal and cannot reduce the bandwidth of a general finite bandwidth signal.

Figure (5.19) illustrates frequency division for a sub-band of the vowel /ε/. Figure (5.19.1) is the amplitude spectrum of the first formant of /ε/. A frequency divided reconstruction by equation (5.20), for n=2, yields the amplitude spectrum, figure (5.19.2). The average instantaneous frequency of the original formant corresponds to the 5th harmonic and division by 2 has caused an average instantaneous frequency shift of 2.5ω (where ω is the spectral line spacing). The resulting spectral lines are at frequencies 0.5ω, 1.5ω, 2.5ω, etc. Any of the original first formant spectral lines below the frequency 2.5ω have been shifted to negative Fourier frequencies in the frequency divided version and therefore fold to +0.5ω, +1.5ω and +2.5ω.

The frequency divided formant is similar to that obtained by frequency shifting and figure (5.19.3) is the amplitude spectrum of the formant reconstructed by

$$\tilde{s}(t) = a(t)\cos\{\int(\omega_i(t) - 2.5\omega).dt)\} \quad \ldots \quad (5.23)$$

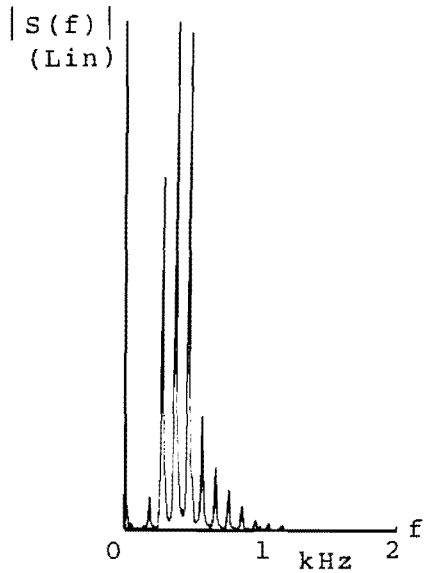


Fig 5.19.1

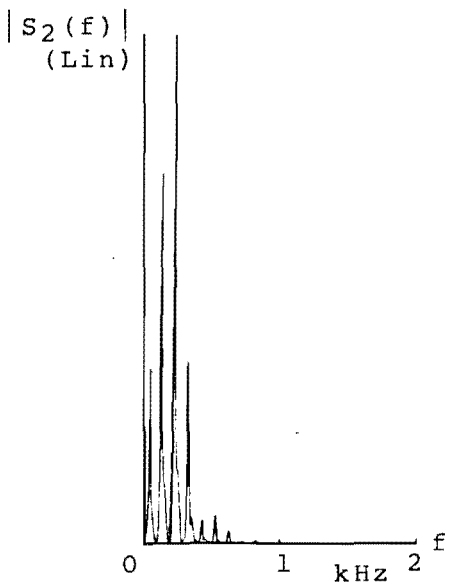


Fig 5.19.2

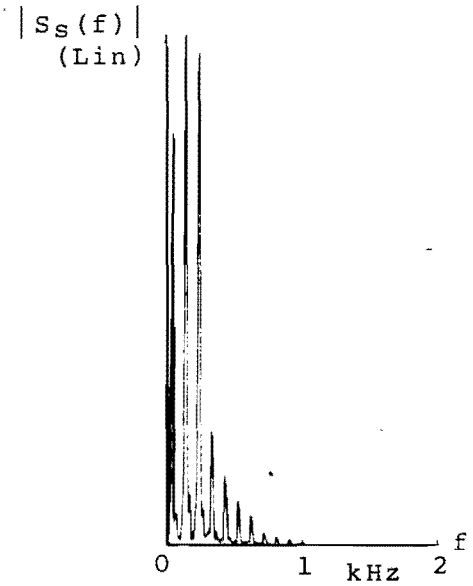


Fig 5.19.3

Fig 5.19 Comparison of Frequency Division and Frequency Shifting

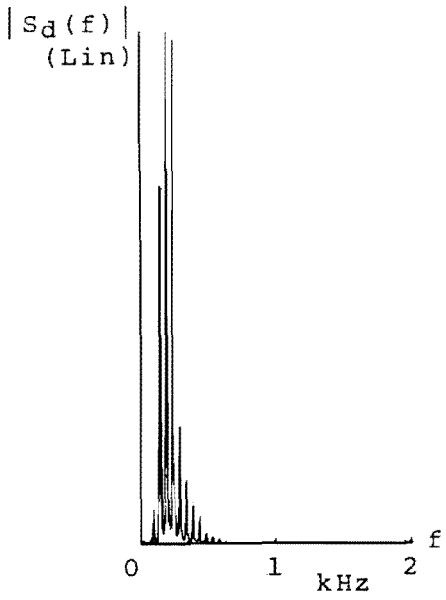


Fig 5.20 True Frequency and Bandwidth Division

A frequency and bandwidth divided version of the formant may be generated from a(t) and $\omega_i(t)$ by linearly interpolating a new sample between each existing instantaneous amplitude and frequency sample, reconstructing by equation (5.20) for n=2, and playing the reconstruction back at the original sampling rate. This time scale alteration restores the correspondence between instantaneous amplitude and frequency fluctuations by restoring the area (phase) enclosed by each instantaneous frequency fluctuation to its value prior to frequency division. The resulting amplitude spectrum, figure (5.20), can be seen to be an accurately scaled version of the original formant.

Unfortunately this technique of distortionless frequency division and bandwidth reduction is not suitable for speech bandwidth reduction as it is exactly equivalent to playing back a recording of speech at half its original rate.

The degree of frequency shifting resulting from a frequency division is dependent on the value of $\overline{\omega_i(t)}$. It has been shown, Section (4.3.1.2), that for a voiced phoneme, $\overline{\omega_i(t)}$ may change from cycle to cycle. This could also happen to a sub-band of speech and frequency division by 2 would produce a formant signal whose centre frequency could step up and down in frequency by 0.5ω, where ω is the spectral line spacing. Problems caused by filtering, transmitting and re-multiplying from such a frequency modulated signal may explain some of the distortions peculiar to frequency division vocoders.

Figure (5.21) is a frequency multiplied reconstruction of the first formant according to

$$\tilde{s}(t) = a(t)\cos\{2x\int\omega_i(t).dt\} \qquad \ldots (5.24)$$

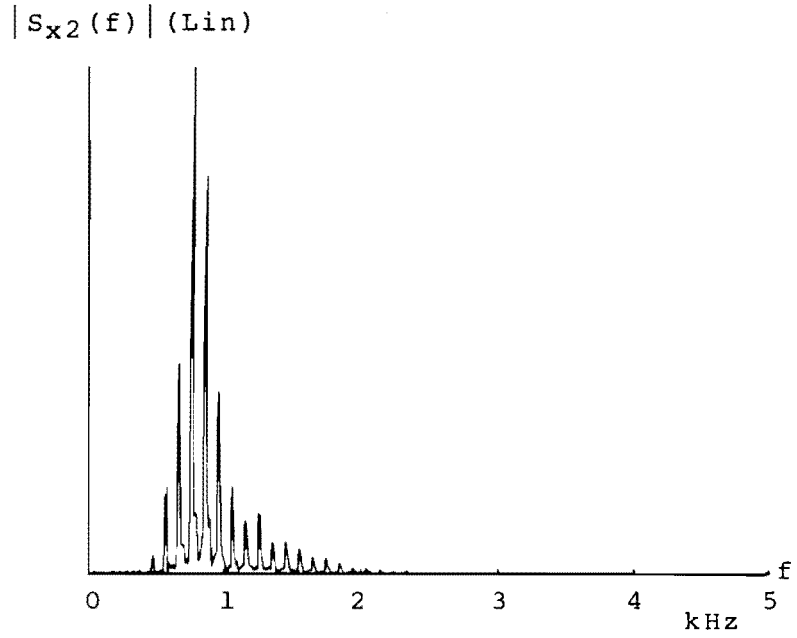


Fig 5.21 Frequency Multiplication (x2)

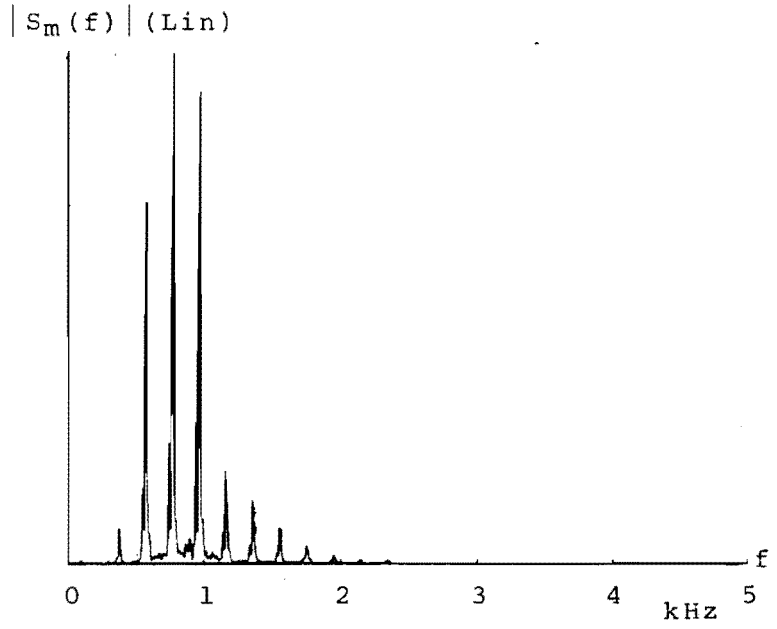


Fig 5.22 Frequency and Bandwidth Multiplication

The average instantaneous frequency is now 10.ω, but once again spectral line spacing is unaffected and bandwidth is slightly increased. An undistorted frequency and bandwidth multiplied first formant may be generated by discarding every second sample of a(t) and $\omega_i(t)$ and reconstructing according to equation (5.24). This process is exactly equivalent to playing the original signal back at twice the rate of recording and the amplitude spectrum of the reconstructed signal is illustrated in figure (5.22).

In order to examine the effects of frequency division and multiplication on a full bandwidth voiced phoneme, it was necessary to avoid "blurring" of the reconstruction (time averaged) amplitude spectrum which could result if average instantaneous frequency were allowed to change from cycle to cycle. For this reason a "periodic vowel" was generated by artificially repeating and joining one cycle of the waveform of the bandpass vowel /ɛ/. Figure (5.23) illustrates the amplitude spectrum, waveform and instantaneous parameters of the periodic vowel which exhibits constant average instantaneous frequency, $\overline{\overline{\omega_i(t)}}=6.\omega$.

Frequency division by 2 shifts average instantaneous frequency to $\overline{\overline{\omega_i(t)}}=3.\omega$ and results in the reconstruction amplitude spectrum, figure (5.24). It can be seen that the vowel formant structure has been obliterated by folded "image" formants and re-analysis and frequency multiplication of the divided signal could not restore the original vowel. Real and "image" formants may only be decoded by reference to the instantaneous frequency waveform used in the generation of the frequency divided signal ($\omega_i(t)/2$). The bandwidth of the frequency divided vowel is similar to that of the vowel spectrum, frequency shifted by -3ω, figure (5.25).

Figure (5.26) is the reconstruction amplitude spectrum after frequency multiplication by 2 and this is comparable

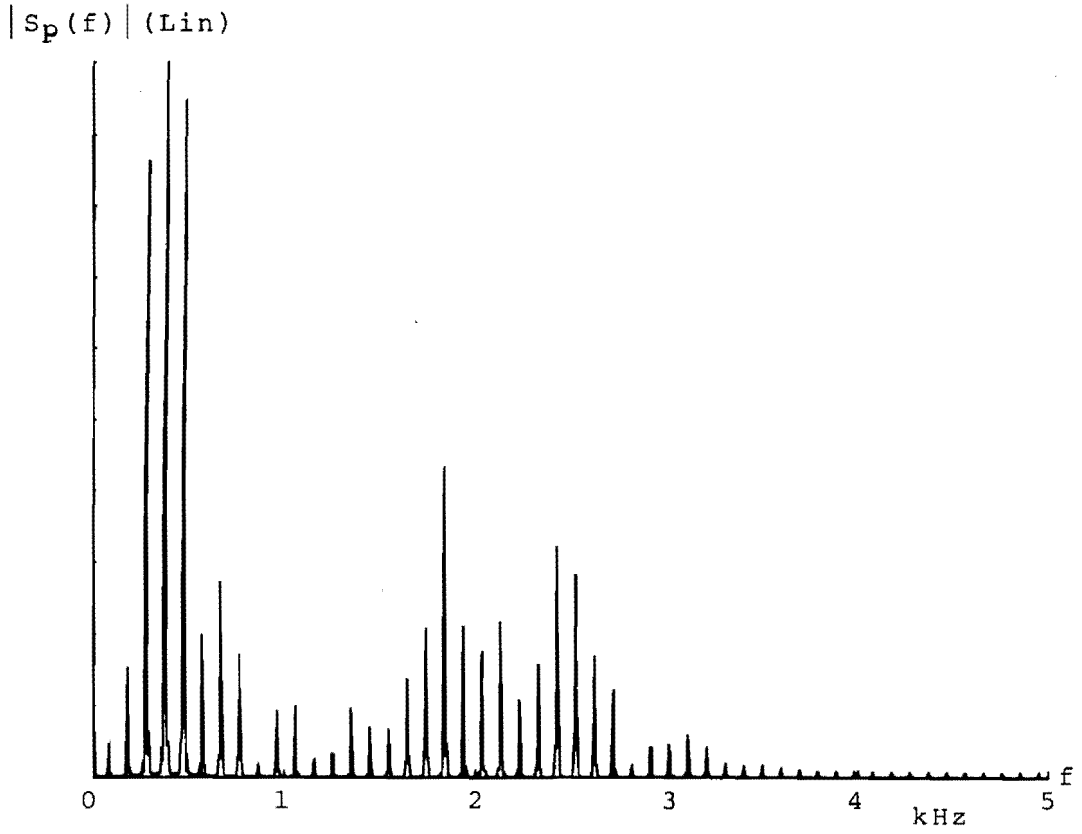


Fig 5.23.1

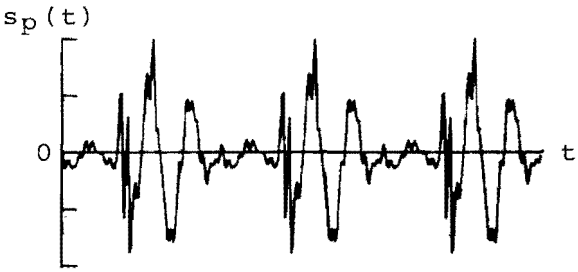


Fig 5.23.2

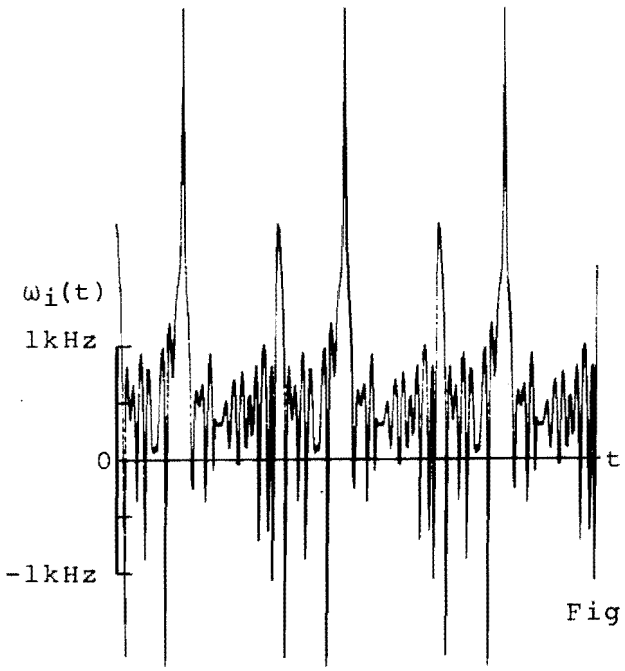


Fig 5.23.3

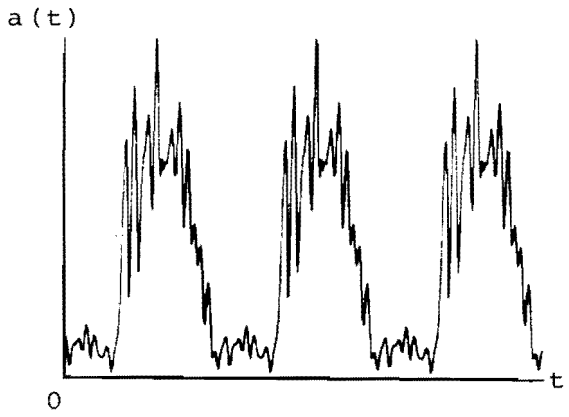


Fig 5.23.4

Fig 5.23 Analysis of "Periodic Vowel"

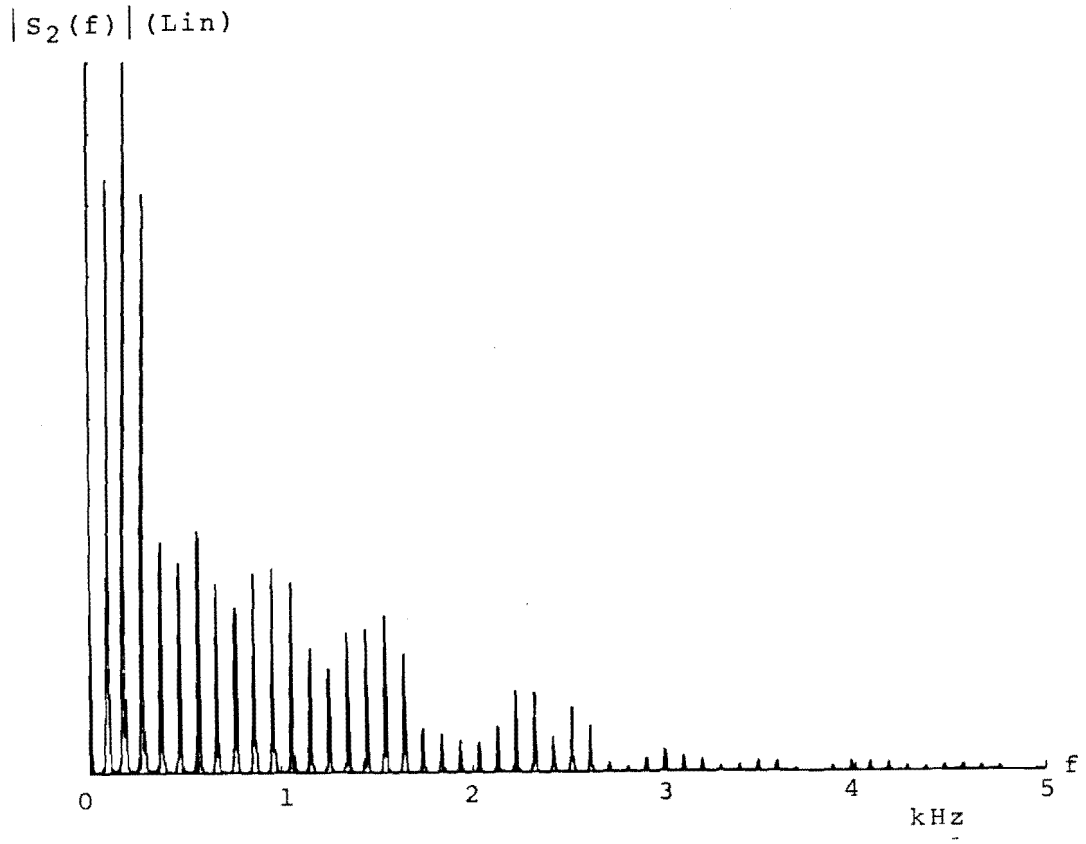


Fig 5.24 Frequency Division (÷2)

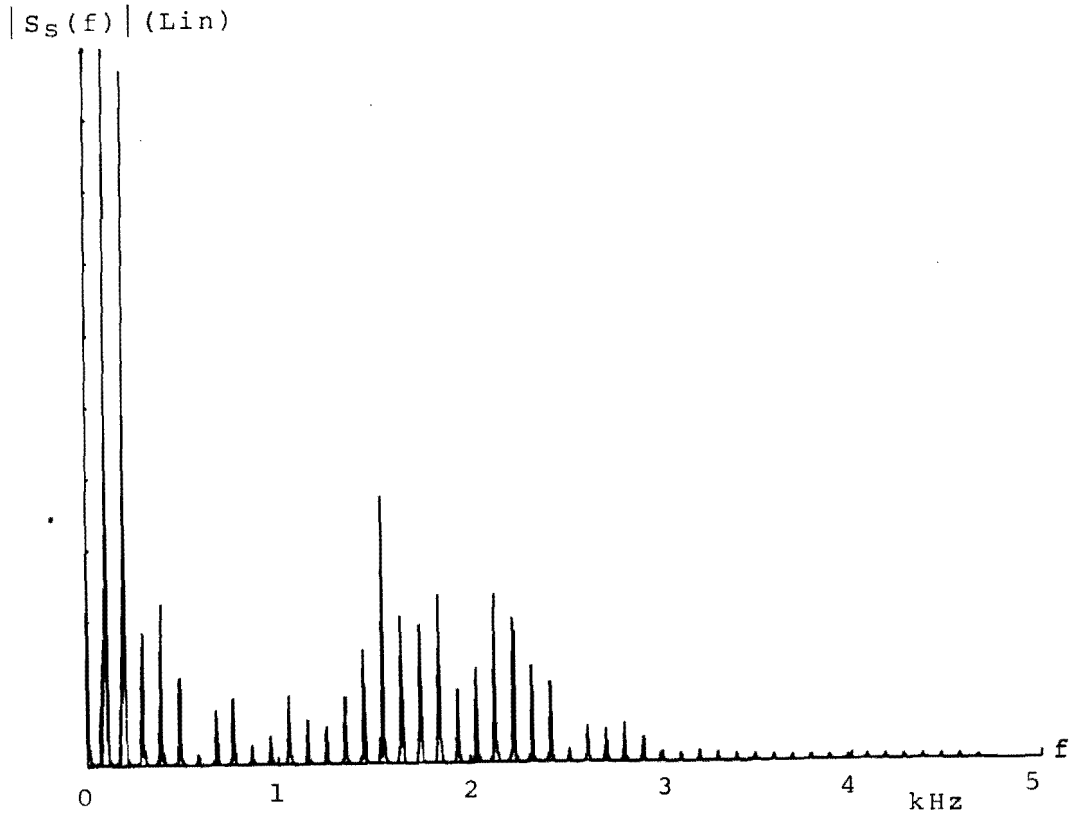


Fig 5.25 Frequency Shift (-3ω)

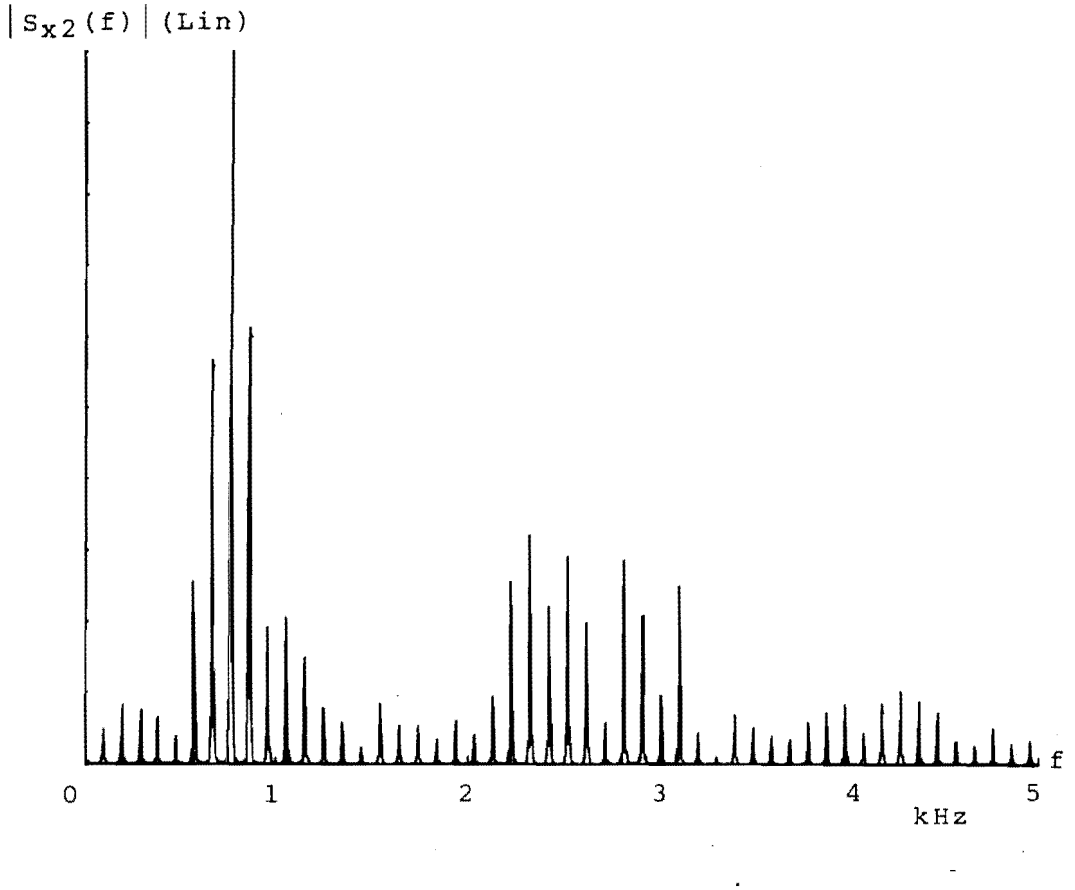


Fig 5.26 Frequency Multiplication (x2)

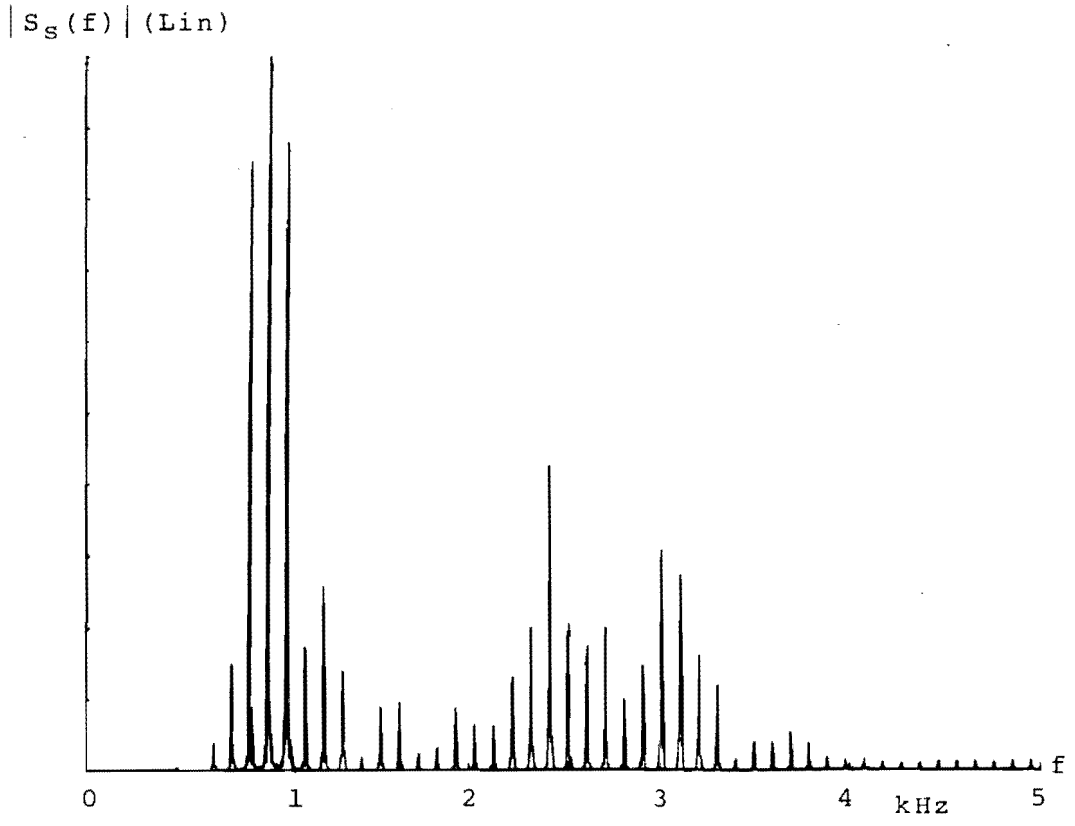


Fig 5.27 Frequency Shift (+6ω)

to the vowel amplitude spectrum shifted by $+6\omega$, figure (5.27).

If the amplitude spectrum of an unvoiced phoneme is assumed to be almost symmetrical about its average instantaneous frequency, frequency division or multiplication by n will result in a frequency shift plus some distortion of the spectral envelope.

(5.6) CONSTANT FREQUENCY RECONSTRUCTION

Reconstructions of speech based on modified versions of its instantaneous parameters

$$\tilde{s}(t) = \tilde{a}(t)\cos\{\int\tilde{\omega}_i(t).dt\} \quad . . . (5.25)$$

have demonstrated the immunity of intelligibility to loss or distortion of the amplitude function. Such robustness of the speech signal suggests that inherent redundancies may be expressed in terms of the amplitude and frequency modulating functions, $a(t)$ and $\omega_i(t)$.

Constant amplitude reconstruction of speech according to

$$\tilde{s}(t) = \cos\{\int\omega_i(t).dt\} \quad . . . (5.26)$$

has been shown to cause symmetricalisation of the signal spectrum around $\overline{\omega_i(t)}$ and, although "image" resonances and increased bandwidth result, $\tilde{s}(t)$ is still highly intelligible. If $s(t)$ is a vowel with spectral envelope, figure (5.28.1) then a frequency shifted constant amplitude reconstruction has a spectral envelope similar to that illustrated in figure (5.28.2)

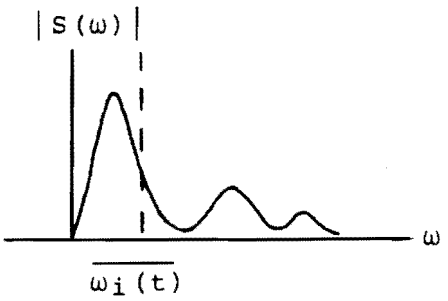


Fig. 5.28.1

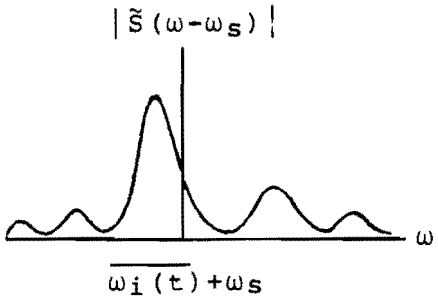


Fig. 5.28.2

Assuming that s(t) is a "periodic vowel" (i.e. one for which the average instantaneous frequency does not change from cycle to cycle), then we may also reconstruct a "constant frequency" version according to

$$g(t) = a(t)\cos\{\omega_m.t\} \qquad \ldots \quad (5.27)$$

where $\omega_m = \overline{\omega_i(t)}$.

The ambiguous nature of information carried by a(t) may be illustrated by analysing the zero structure of $a^2(t)$. If $\Psi(t)$ is the analytic signal defined by a(t) and $\omega_i(t)$, then noting the relation $a^2(t)=\Psi(t).\Psi^*(t)$, it can be seen that the square of the instantaneous amplitude function possesses all of the zeros of $\Psi(t)$ and $\Psi^*(t)$. The zeros of $a^2(t)$ must therefore all occur in complex conjugate pairs. Factorisation of $a^2(t)$ provides the zero locations of $\Psi(z)$ only if additional information is provided to resolve each of them into the LHP or UHP.

It has been shown (Section (4.3.1), figure (4.4.5)) that a(t) may possess a resonance structure based on the formants of s(t), and g(t) will therefore exhibit a frequency shifted spectral envelope similar to that illustrated in figure (5.28.3).

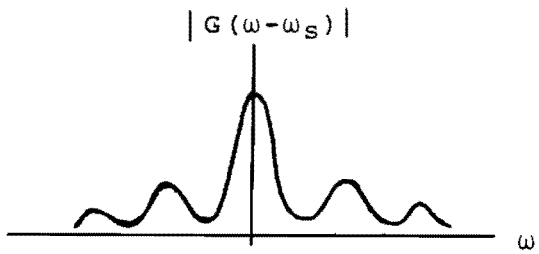


Fig. 5.28.3

This constant frequency reconstruction places the first formant frequency at exactly $\overline{\omega_i(t)}$, and amplitude modulation distributes a(t) symmetrically around the carrier. The "image" and real formants are therefore of the same magnitude and are indistinguishable.

Although oversimplified, this example illustrates the relationship between information conveyed by the amplitude and frequency modulating functions, and suggests that a constant frequency reconstruction of speech may be at least partially intelligible.

In practise, straightforward constant frequency reconstruction of speech is not practical as average instantaneous frequency must be allowed to change from phoneme to phoneme and, in the case of a voiced fricative, within one cycle of a phoneme. An approximation to constant frequency voice may be constructed, however, by

$$\tilde{s}(t) = a(t)\cos\{\int\tilde{\omega}_i(t).dt\} \qquad \ldots (5.28)$$

where $\tilde{\omega}_i(t)$ is a lowpass version of $\omega_i(t)$. A passband of "a few hundred Hz" ensures the removal of most instantaneous frequency fluctuations without greatly distorting the changes of $\overline{\omega_i(t)}$. As a lowpass operation, this type of filtering has the advantage of not affecting the area enclosed by the instantaneous frequency curve, and therefore imparts no frequency shift to $\tilde{s}(t)$.

The minimum distortion of the "shape" of $\overline{\omega_i(t)}$ changes is ensured by using a linear phase lowpass filter. For this reason all lowpass instantaneous frequency waveforms have been generated using a truncated FIR implementation of the cosine rolloff impulse response.

$$h(t)=2.f_c.\mathrm{sinc}(2f_ct)\frac{\cos(2\pi\rho f_ct)}{1-16\rho^2f_c^2t^2} \quad . \ . \ . \quad (5.29)$$

where f_c is the lowpass cutoff frequency and ρ is the rolloff factor. This impulse response and the corresponding frequency response are illustrated for various values of ρ in figure (5.29) (Ref. 67). For most practical situations, it has been found that h(t) may be truncated on both sides of t=0 when the value of the maximum of $|h(|t|)|$ between an adjacent pair of zero crossings is less than h(0)/100.

An initial investigation into the effects of lowpass filtering of instantaneous frequency waveforms and on the resulting reconstructions was performed using a frequency shifted version of the vowel /ɛ/. The amplitude spectrum, time waveform and instantaneous parameters of the vowel are those illustrated previously in figure (5.10).

The principal rate of fluctuation of instantaneous frequency for this vowel has been shown to correspond to the difference frequency between the first and second formants. As this frequency is approximately 1,500 Hz,

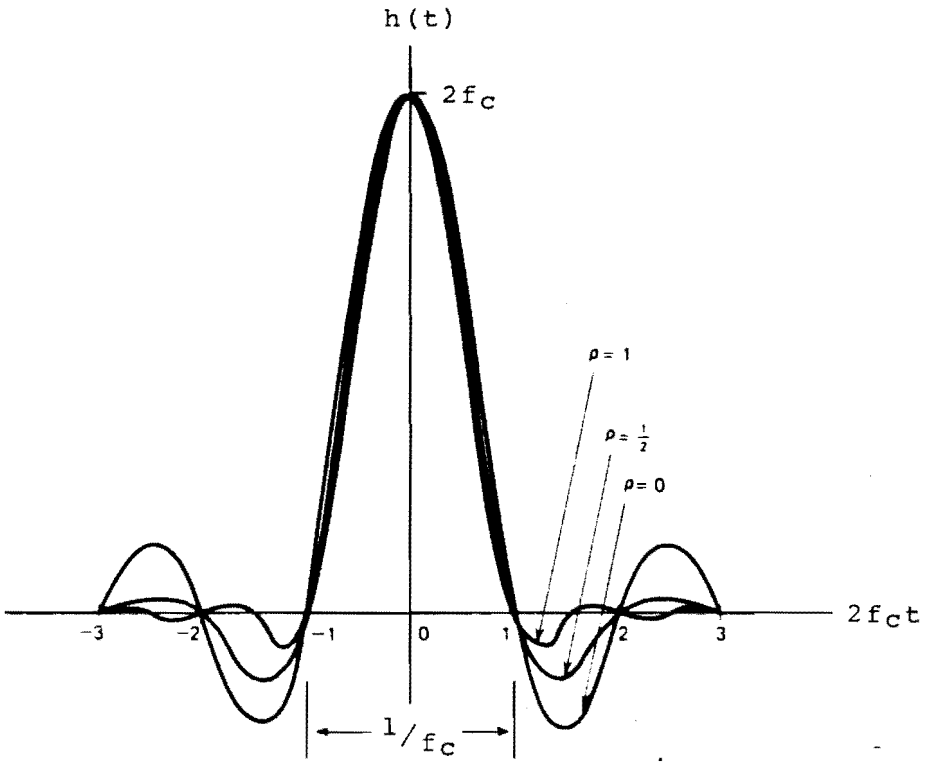


Fig 5.29.1

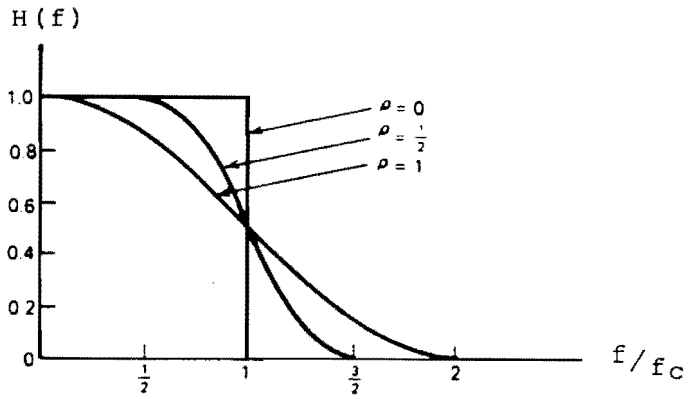


Fig 5.29.2

Fig 5.29 Possible Impulse and Frequency Responses for Lowpass Cosine Rolloff Filter

lowpass filtering of instantaneous frequency to 1,000 Hz should remove second formant information, but maintain $\overline{\omega_i(t)}$. Figure (5.30.1) is $\omega_i(t)$ lowpass filtered to f_c=1,000 Hz, ρ=0.3 and figures (5.30.2) and (5.30.3) are the corresponding filter impulse and frequency responses.

Comparison of figures (5.10.3) and (5.30.1) reveals that lowpass filtering has caused a sizeable reduction in the amplitude of instantaneous frequency fluctuations. If it is assumed that the most "significant" fluctuations of unbandlimited instantaneous frequency, about $\overline{\omega_i(t)}$, can be approximated by Dirac delta functions of area $\pm$ ½ cycle ($\pm\pi$ radians), then the amplitude of lowpass instantaneous frequency fluctuations, about $\overline{\omega_i(t)}$, may be estimated for a given f_c.

The response, r(t), of the lowpass filter to one of these impulses (at time t=0) is given by

$$\begin{aligned} r(t) &= h(t) * (\pm\tfrac{1}{2})\,\delta(t) \\ &= (\pm\tfrac{1}{2})\,h(t) \end{aligned} \qquad \ldots \quad (5.30)$$

The maximum amplitude fluctuation of lowpass instantaneous frequency, $r_{max}(t)$, will occur when n unbandlimited impulses, of the same polarity, occur within a period of approximately ½f_c seconds. As h(t) has an effective width of $1/f_c$ seconds, the n impulses may be assumed to have occurred simultaneously and the resulting lowpass function is

$$\begin{aligned} r_{max}(t) &= h(t) * (\pm n/2)\,\delta(t) \\ &= (\pm n/2)\,h(t) \end{aligned} \qquad \ldots \quad (5.31)$$

Unity gain in the lowpass passband requires $h(0)=2f_c$, and as this is the maximum of h(t), the greatest possible deviation of lowpass instantaneous frequency about $\overline{\omega_i(t)}$ is

$$\begin{aligned} r_{max}(t) &= (\pm n/2).2f_c \\ &= \pm n.f_c \end{aligned} \qquad \ldots \quad (5.32)$$

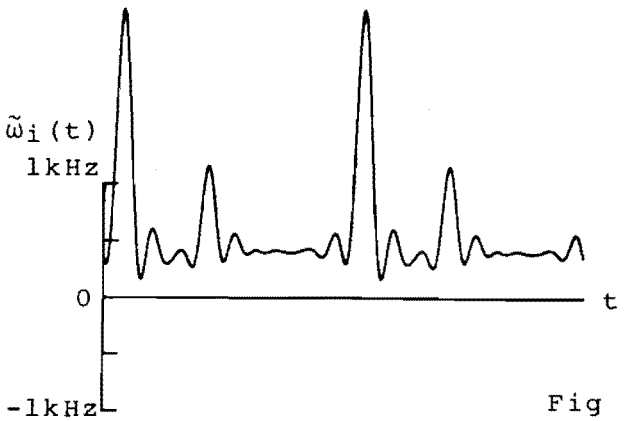


Fig 5.30.1

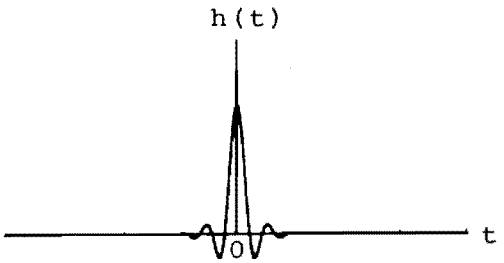


Fig 5.30.2

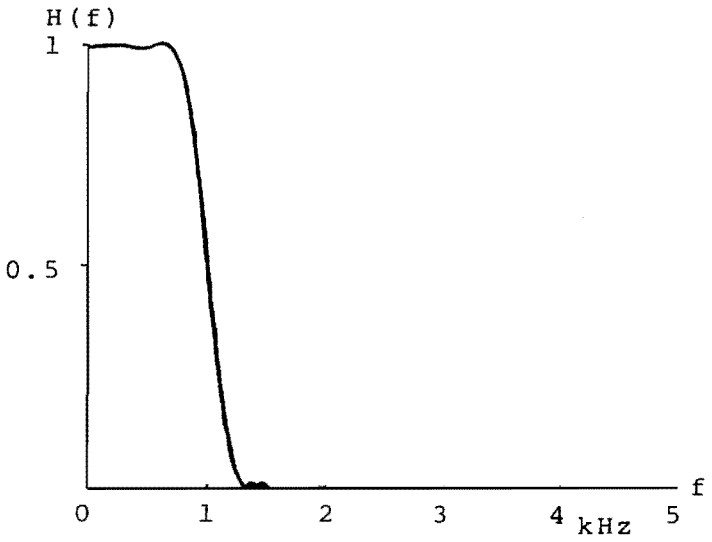


Fig 5.30.3

Fig 5.30 Illustration of Lowpass Instantaneous Frequency

Detailed analysis of this vowel in Section (4.3.1.1) has shown that the negative instantaneous frequency excursions of figure (5.10.3) are due to single LHP analytic signal zeros and, as such, they each enclose an area of $-\frac{1}{2}$ cycle. If they were sufficiently "impuse like", these dips could cause a lowpass fluctuation of amplitude $r(t)=-f_c$, which corresponds to an absolute instantaneous frequency of approximately $\overline{\omega_i(t)}/2\pi-f_c$. In the above case, $\overline{\omega_i(t)}/2\pi\simeq 600$ Hz and $f_c=1{,}000$ Hz, so the minimum possible instantaneous frequency is -400 Hz. The fact that the instantaneous frequency curve of figure (5.30.1) never becomes negative suggests that the unbandlimited instantaneous frequency dips are not good approximations of Dirac impulses.

The single positive instantaneous frequency spike per cycle of figure (5.10.3) is caused by the approximate temporal superposition of one UHP analytic signal zero (due to the major harmonics of the first formant) and a second UHP zero which has been generated by LHP to UHP conversion (Section (4.3.1.1)). The result is an impulse like function of instantaneous frequency, with area +1 cycle about $\overline{\omega_i(t)}/2\pi$, which could cause a lowpass instantaneous frequency fluctuation of amplitude $r(t)= +2f_c$. In terms of absolute instantaneous frequency, $\overline{\omega_i(t)}/2\pi +2f_c=2{,}600$ Hz for the above case. The maximum of the large positive instantaneous frequency fluctuation in figure (5.30.1) is +2,540 Hz, indicating that the two UHP zero instantaneous frequency spike is a good approximation to a Dirac impulse.

Performing a frequency shifted <u>constant amplitude</u> reconstruction with the 1,000 Hz lowpass instantaneous frequency waveform yields the amplitude spectrum, figure (5.31). As expected, all second formant information has been obliterated and bandwidth is asymmetrical about $\overline{\omega_i(t)}$. The bandwidth above $\overline{\omega_i(t)}$ is governed by the maximum amplitude fluctuation of lowpass instantaneous frequency, $r(t)=+2f_c$,

but there are also strong components at $\overline{\omega_i(t)}/2\pi \pm f_c$. These are due to narrow band frequency modulation at frequency f_c caused by the decaying quasi-periodic tails of h(t) which have frequency f_c when $\rho<0.5$, figure (5.29). The amplitude spectrum resulting from a full bandwidth instantaneous frequency, frequency shifted, constant amplitude reconstruction of the vowel is shown in figure (5.32).

Reconstructing the vowel with 1,000 Hz lowpass instantaneous frequency and the full bandwidth amplitude function

$$\tilde{s}(t) = a(t)\cos\{\int \tilde{\omega}_i(t).dt\} \qquad \ldots (5.33)$$

yields the amplitude spectrum, figure (5.33). Comparison with figure (5.31) confirms that reintroduction of the amplitude function has restored the formant structure. Simple amplitude modulation of the constant carrier frequency $\overline{\omega_i(t)}$ would result in a symmetrical formant structure, but the remaining assymetry of $\tilde{\omega}_i(t)$ ensures that "image" formants are of reduced amplitude.

(5.7) RECONSTRUCTION FROM BANDLIMITED PARAMETERS

The result of this initial test suggests that the approximate constant frequency reconstruction of speech could be as intelligible as constant amplitude voice. If it is found that intelligible speech results from reconstructions in which both the amplitude and frequency modulating functions have been reduced in bandwidth, the possibility exists for the development of a bandwidth efficient speech transmission system. In order to obtain a feel for the relationship between reconstructed speech quality and instantaneous parameter bandwidth, the following analyses were performed on three individual phonemes.

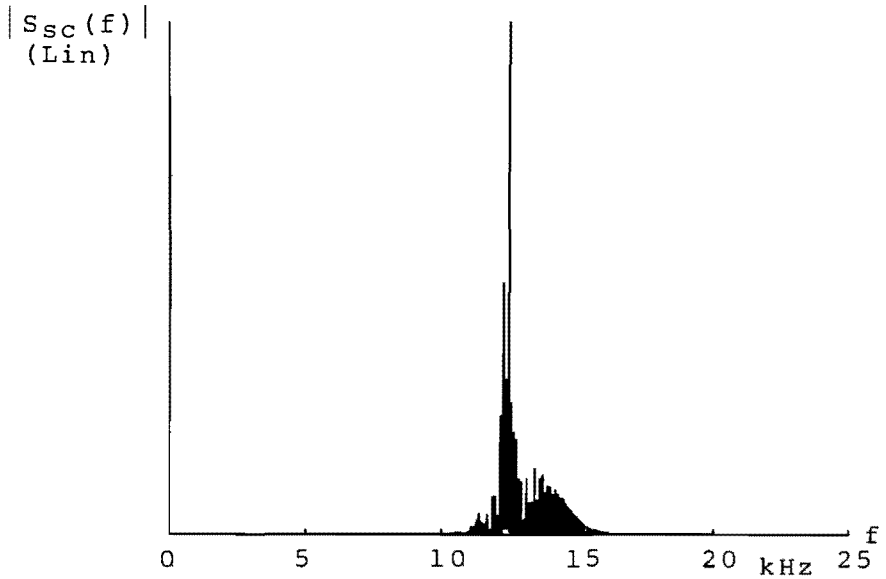


Fig 5.31 Frequency Shifted, Constant Amplitude Lowpass Instantaneous Frequency Reconstruction

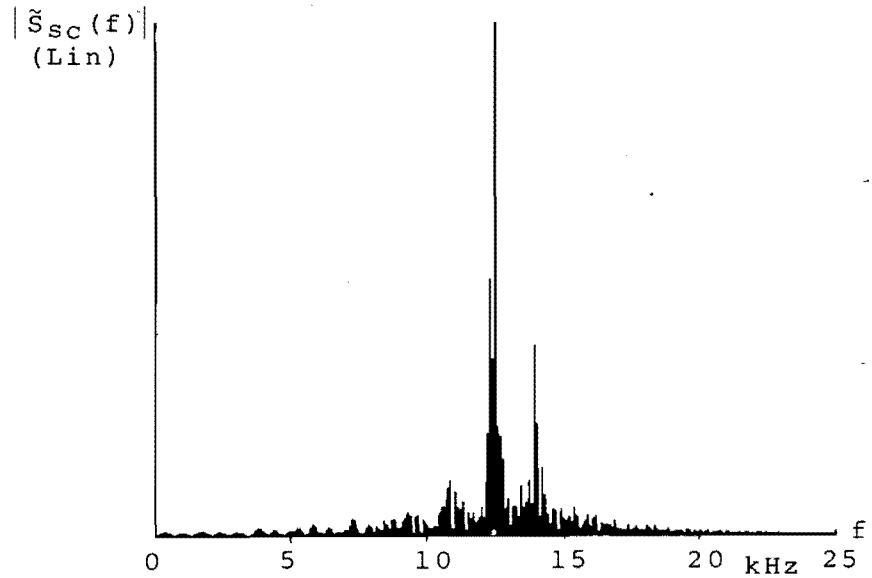


Fig 5.32 Frequency Shifted, Constant Amplitude Full Bandwidth Instantaneous Frequency Reconstruction

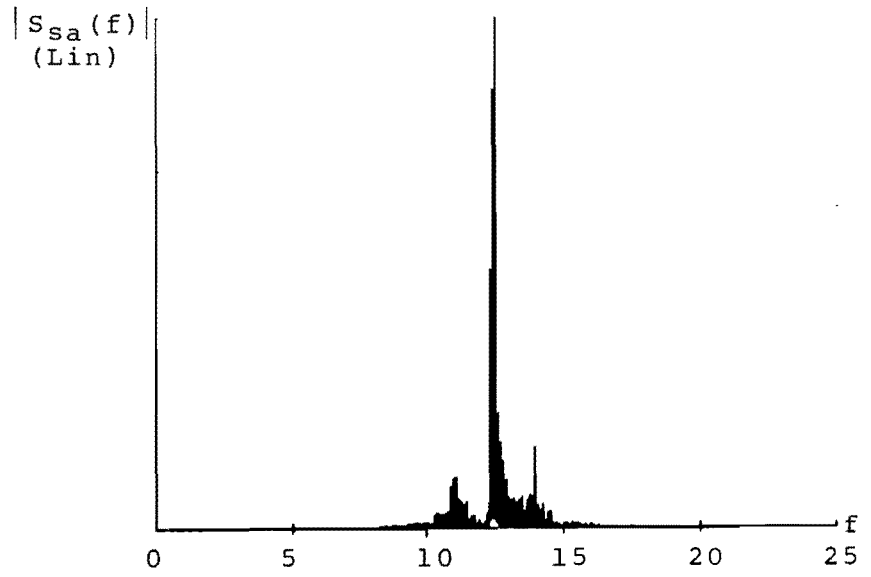


Fig 5.33 Frequency Shifted, Full Bandwidth Instantaneous Amplitude Lowpass Instantaneous Frequency Reconstruction

(5.7.1) VOWEL

In this analysis, /ɛ/ is reconstructed with combinations of a(t) and $\omega_i(t)$ lowpass filtered to 2,200 Hz, 1,000 Hz and 500 Hz. For each case, appropriate waveforms and amplitude spectra may be compared with the amplitude spectrum, time waveform, instantaneous functions and vector locus of the undistorted vowel, figure (5.34).

Each of the following three sets of reconstructions shows the result of reconstructing with a particular lowpass instantaneous frequency waveform and either full bandwidth instantaneous amplitude, or the amplitude function lowpass filtered to 500 Hz. Lowpass filtering a(t) to 500 Hz, figure (5.35), effectively removes any high frequency formant information, but maintains the waveforms basic shape.

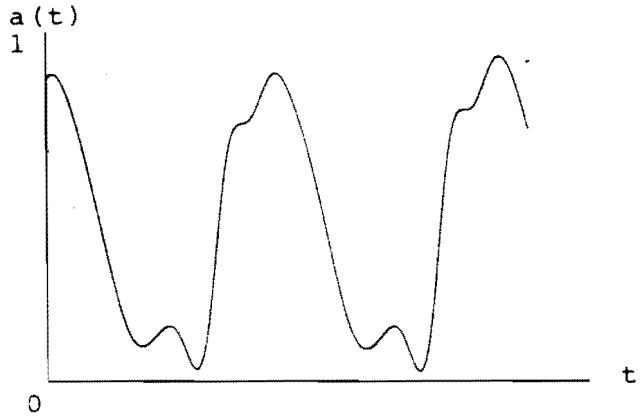


Fig. 5.35 Lowpass Amplitude Function

The first pair of reconstructions are for $\omega_i(t)$ lowpass filtered to 2,200 Hz, figure (5.36.1). As the difference frequency between the first and third formants is only 2,000 Hz, this filtering is expected to have little effect on formant information conveyed by $\omega_i(t)$.

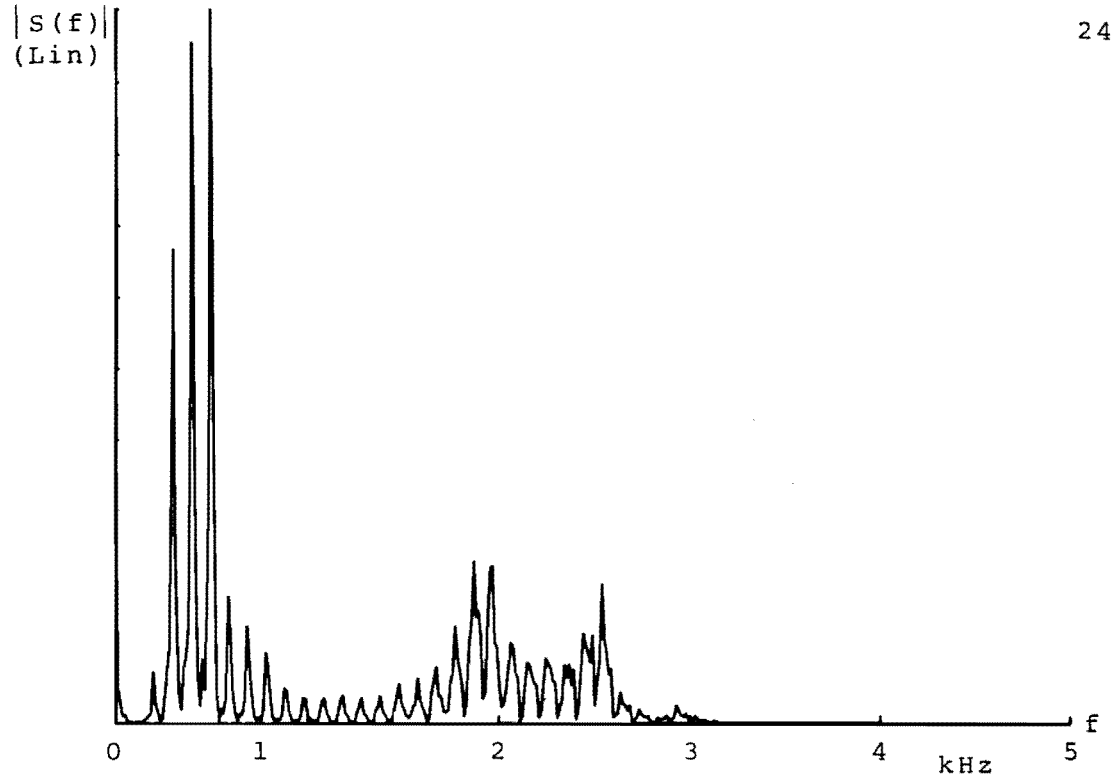


Fig 5.34.1

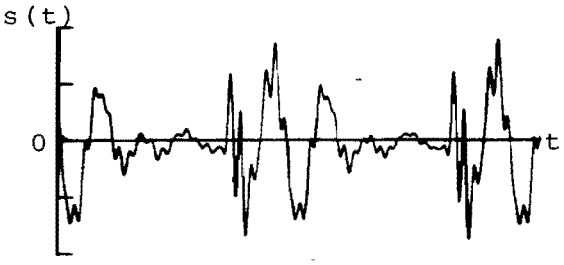


Fig 5.34.2

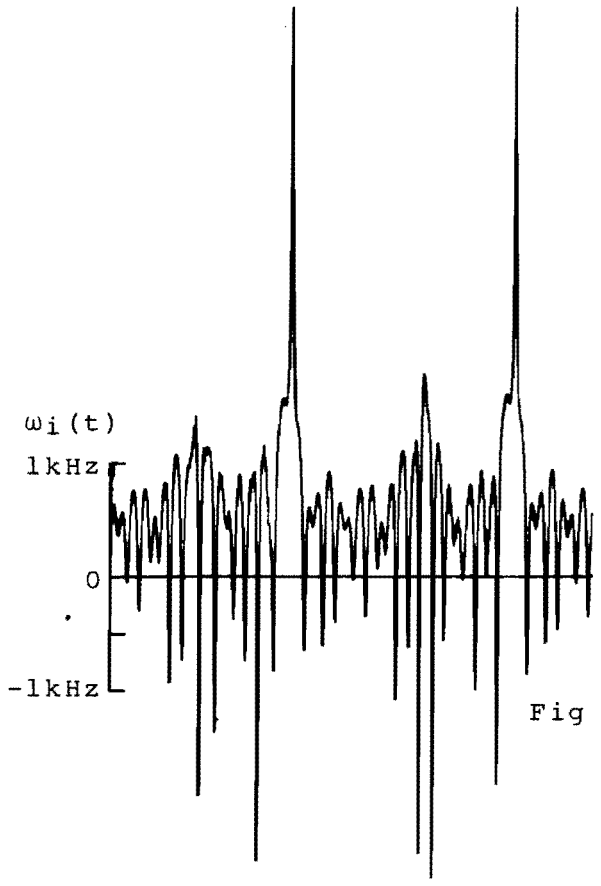


Fig 5.34.3

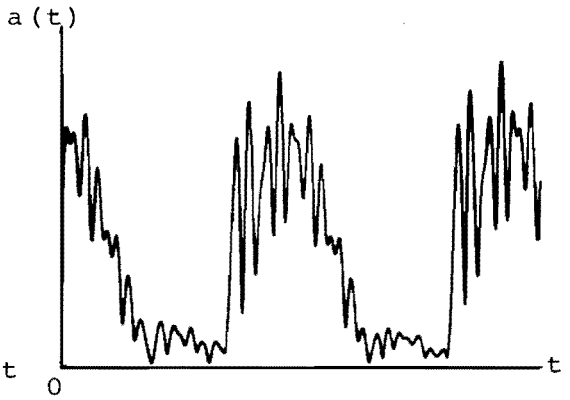


Fig 5.34.4

Figs 5.34.1-5.34.4 Reference Vowel

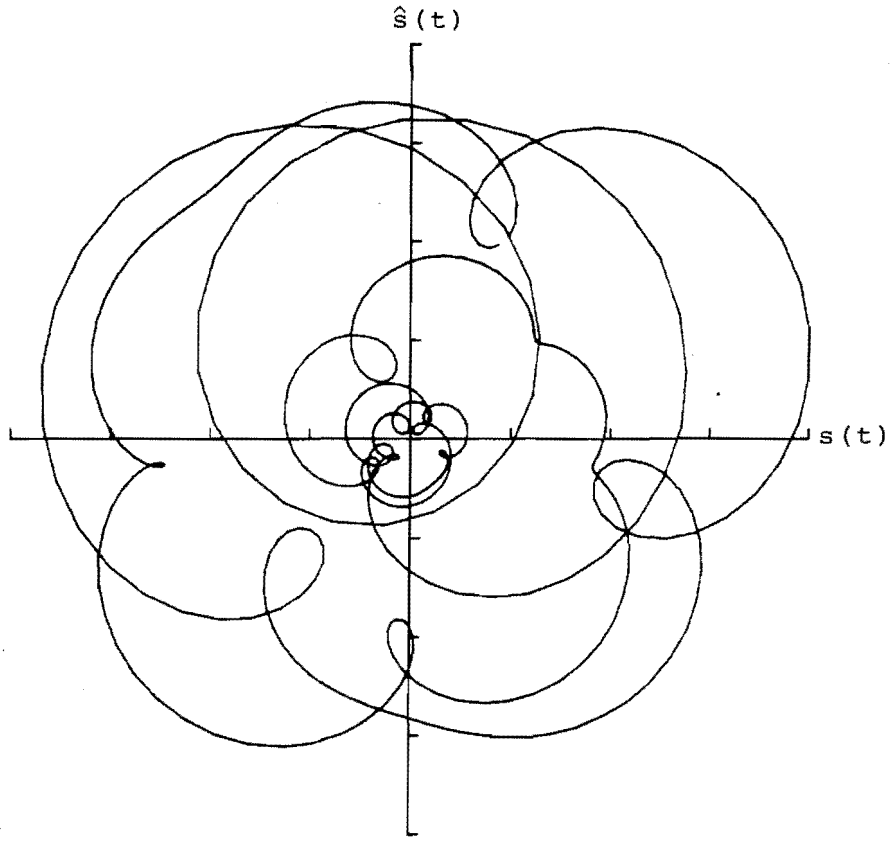


Fig 5.34.5 Reference Vowel

Reconstructing with full bandwidth instantaneous amplitude results in the waveform and amplitude spectrum, figures (5.36.2) and (5.36.3). These are very similar to the original. Reconstruction with the lowpass version of a(t) produces the waveform, figure (5.36.4), and the spectrum, figure (5.36.5), which resembles that of a constant amplitude reconstruction of /ɛ/. Notable features of this spectrum are the presence of an "image" resonance between the first and second formants, the reduced amplitudes of second and third formants and the increased bandwidth.

The second set of reconstructions are with $\omega_i(t)$ filtered to 1,000 Hz, figure (5.37.1). The waveform has been shown to exhibit a maximum frequency deviation of 2,000 Hz above $\overline{\omega_i(t)}/2\pi$, and is sufficiently lowpass to contain no formant frequency difference information.

The full bandwidth a(t) reconstruction produces the waveform, amplitude spectrum and vector locus, figures (5.37.2), (5.37.3) and (5.37.4). In this case, the amplitude spectrum still resembles the original, but the formant structure is distorted by "image" resonances. The vector locus has been grossly distorted by the exclusion of negative instantaneous frequencies.

Reconstruction with lowpass instantaneous amplitude generates the waveform and amplitude spectrum, figures (5.37.5) and (5.37.6). The spectrum resembles that of the constant amplitude, approximate constant frequency reconstruction, figure (5.31). There is a spectral peak at $\overline{\omega_i(t)}/2\pi+f_c$=1,600 Hz and the spectrum is bandlimited to just over $\overline{\omega_i(t)}/2\pi+2f_c$=2,600 Hz. Second and third formant information appears to have been lost in this reconstruction.

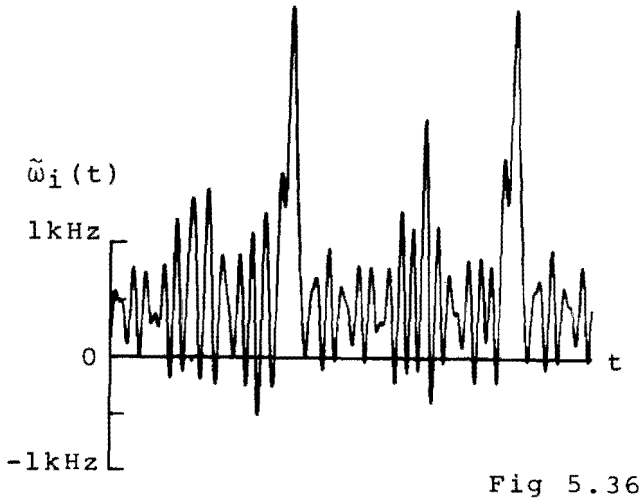


Fig 5.36.1

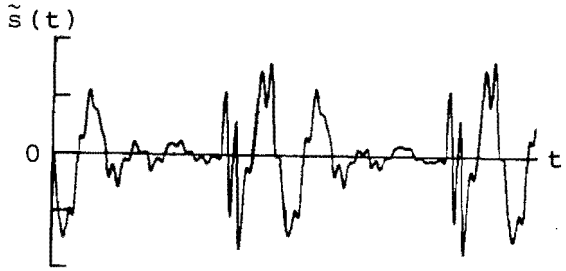


Fig 5.36.2

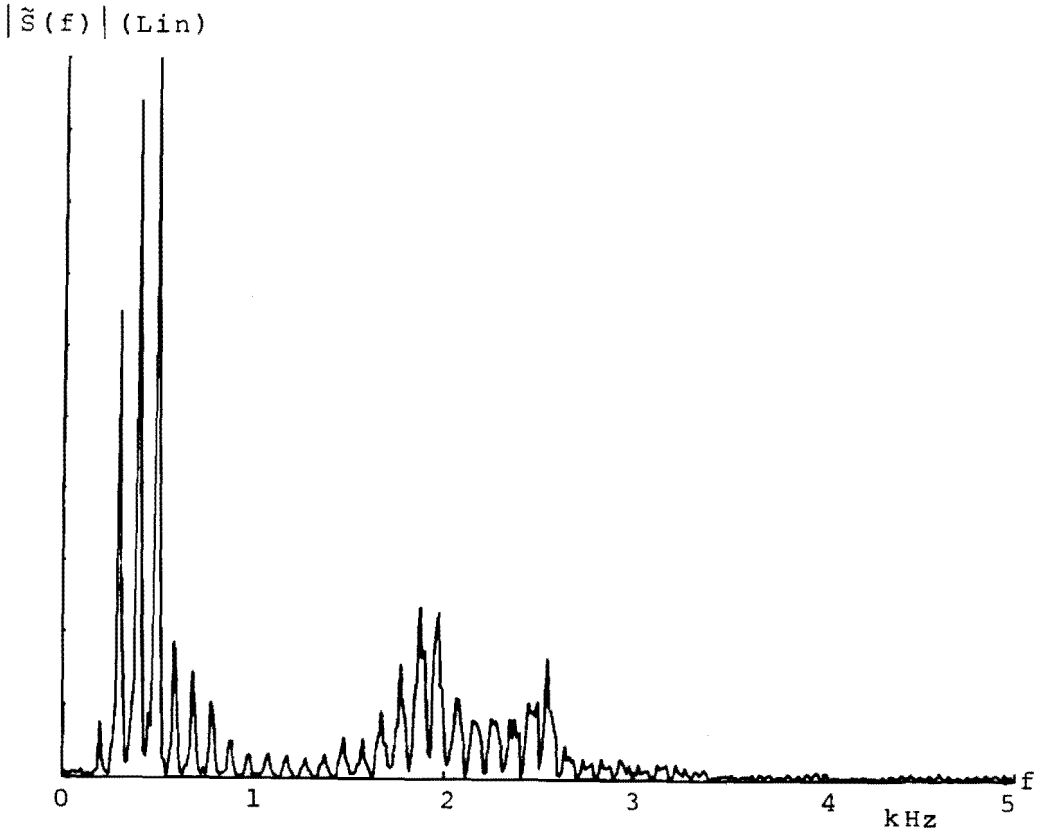


Fig 5.36.3

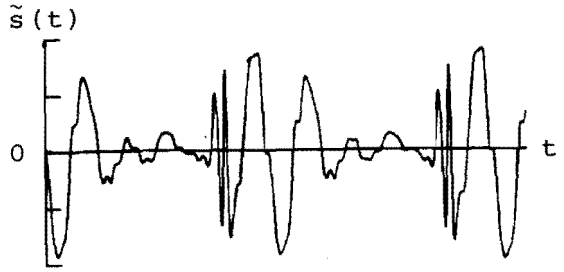


Fig 5.36.4

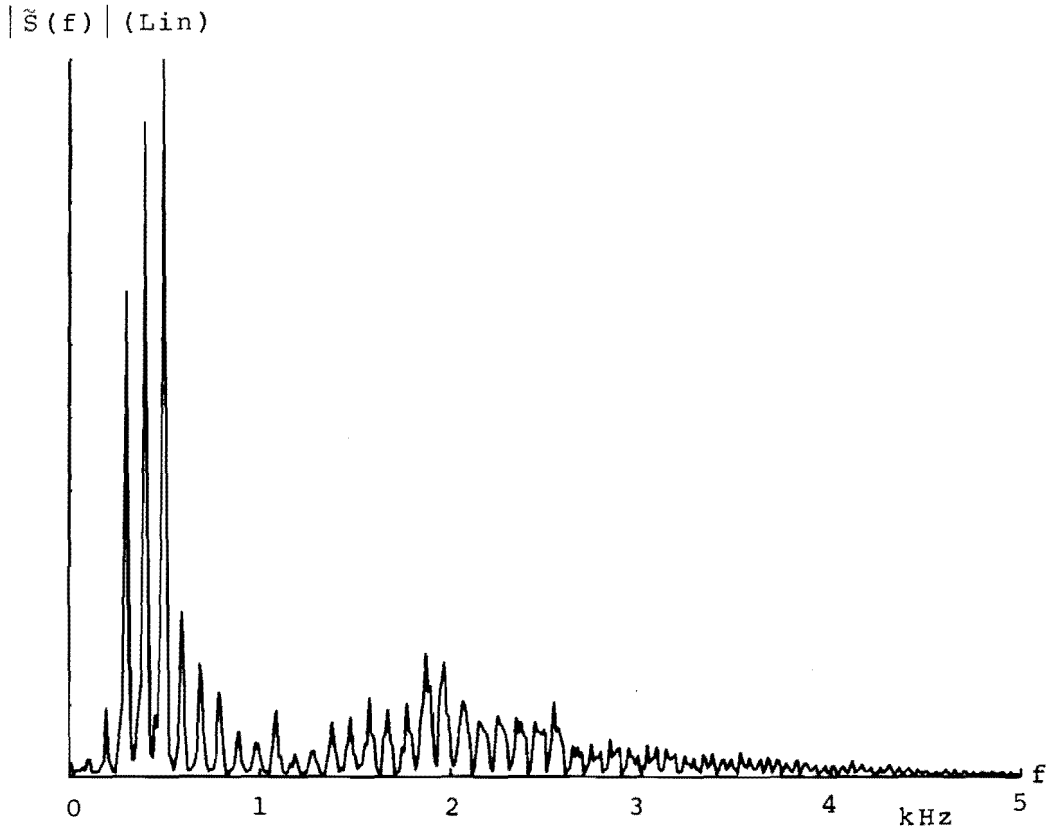


Fig 5.36.5

Fig 5.36 Lowpass Instantaneous Parameter Reconstructions
($\tilde{\omega}_i(t)$ @ 2200Hz)

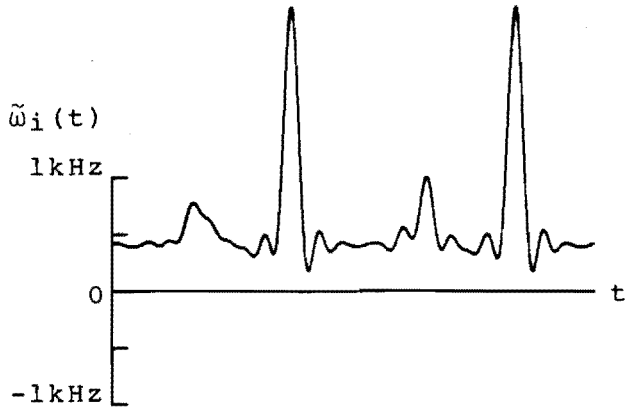


Fig 5.37.1

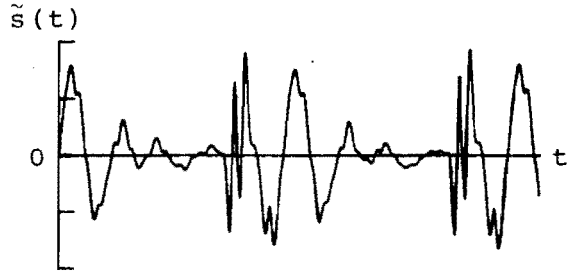


Fig 5.37.2

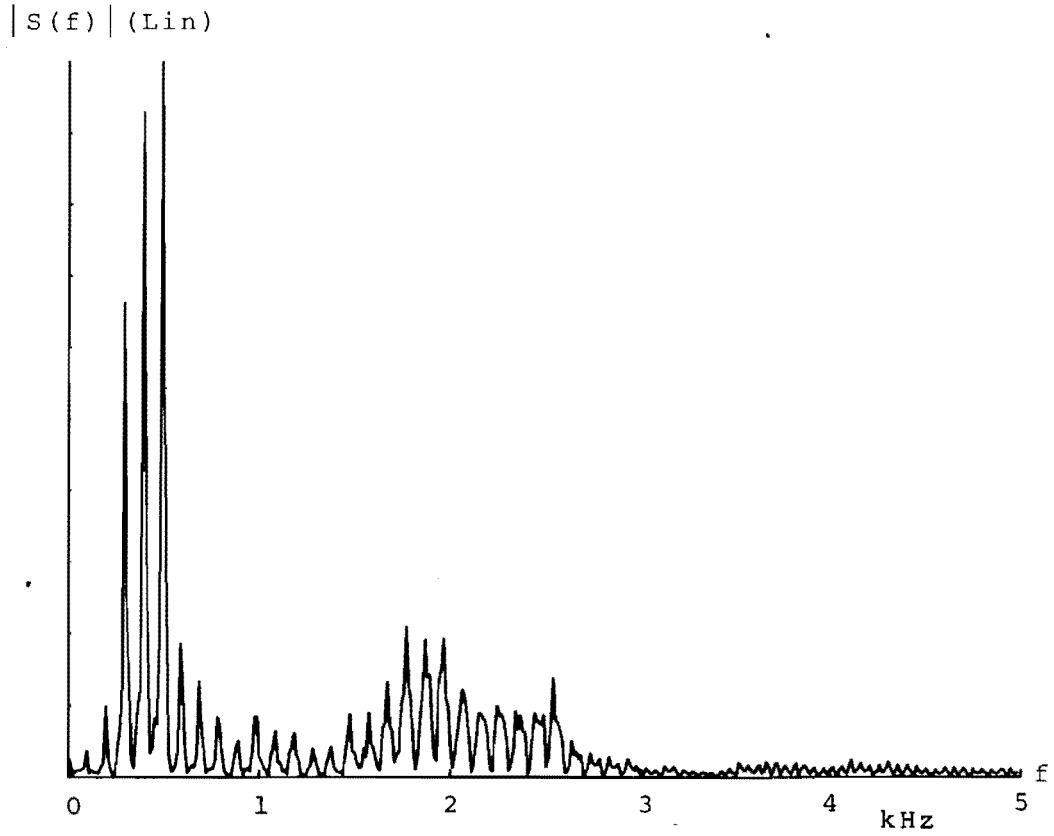


Fig 5.37.3

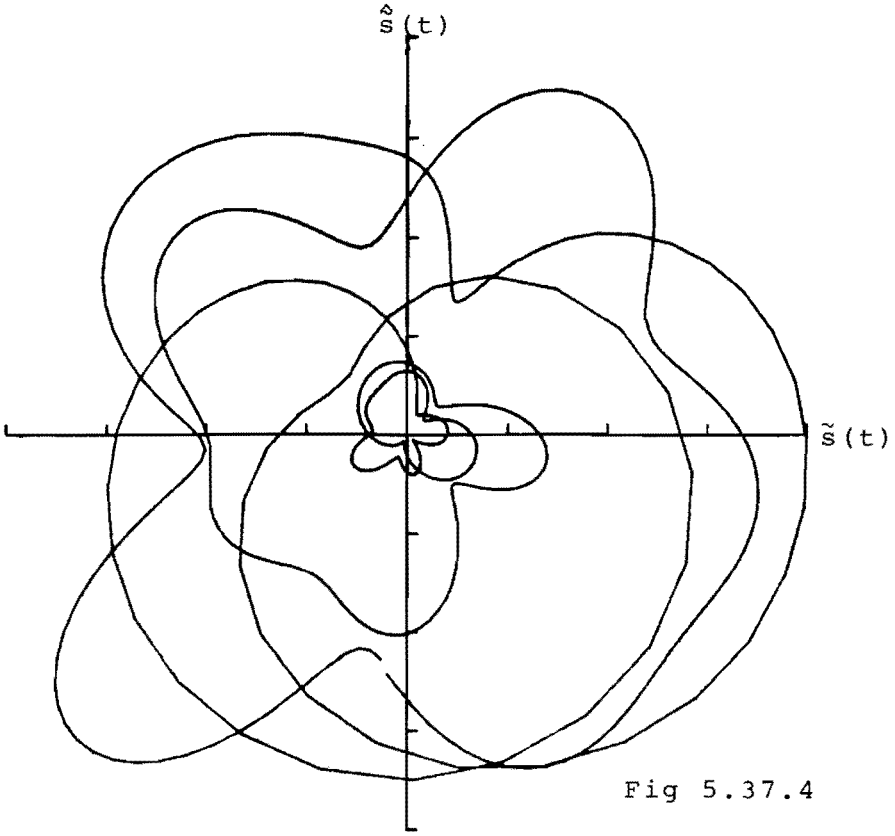


Fig 5.37.4

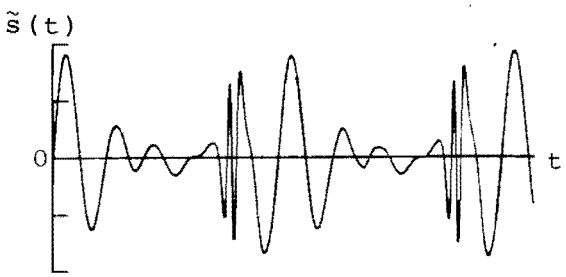


Fig 5.37.5

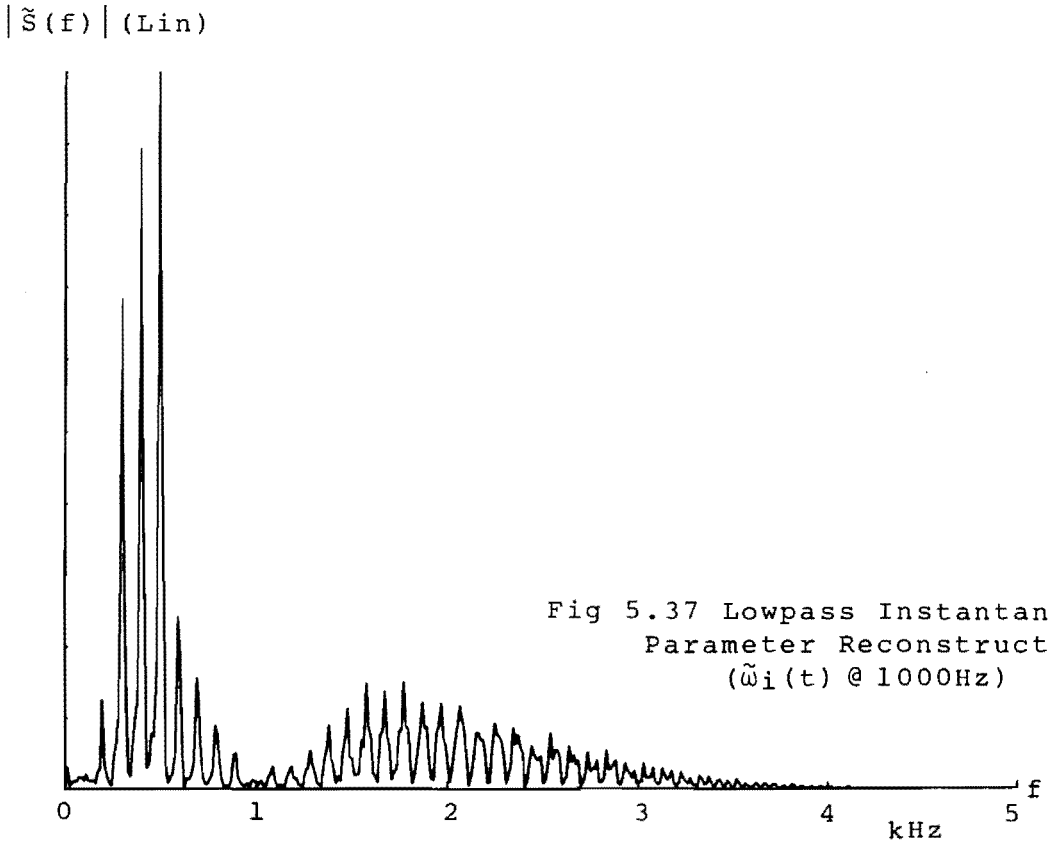


Fig 5.37 Lowpass Instantaneous Parameter Reconstructions ($\tilde{\omega}_i(t)$ @ 1000Hz)

Fig 5.37.6

The time waveform, figure (5.37.5), exhibits a 180° reconstruction phase shift and an obvious loss of high frequency information when compared with the original, figure (5.34.2), but several zero crossing positions are almost exactly preserved. This will be the case as long as fluctuations of the lowpass instantaneous frequency waveform can reach sufficient amplitude. In the limit as instantaneous amplitude and frequency bandwidths are reduced to zero, the zero crossing positions of the reconstructed waveform will approach those of a sinusoid at frequency $\overline{\omega_i(t)}/2\pi$.

The final set of reconstructions involve $\omega_i(t)$ lowpass filtered to f_c=500 Hz, figure (5.38.1). This reduces the maximum instantaneous frequency fluctuation to approximately $2.f_c$=1,000 Hz and should introduce strong spectral components into the reconstructions at $\overline{\omega_i(t)}/2\pi$+500 Hz.

The amplitude spectra generated by full bandwidth instantaneous amplitude reconstruction, figure (5.38.2) and lowpass instantaneous amplitude reconstruction, figure (5.38.3), are very similar and this suggests that the frequency modulating function dominates spectral shaping. In both cases there is a spectral peak at $\overline{\omega_i(t)}/2\pi$+500=1,100 Hz and little bandwidth above $\overline{\omega_i(t)}/2\pi$+1,000=1,600 Hz.

The trend indicated by the above reconstructions is that to ensure "reasonable" reconstruction of a vowel waveform, either the amplitude or frequency modulating function must be retained at bandwidth equal to or greater than the highest formant difference frequency. This rule requires modiciation, however, as the reconstruction from full bandwidth instantaneous amplitude and instantaneous frequency lowpass filtered to 500 Hz did not produce a good vowel reconstruction. It would appear, therefore, that there is a minimum allowable instantaneous frequency bandwidth for good vowel reconstruction.

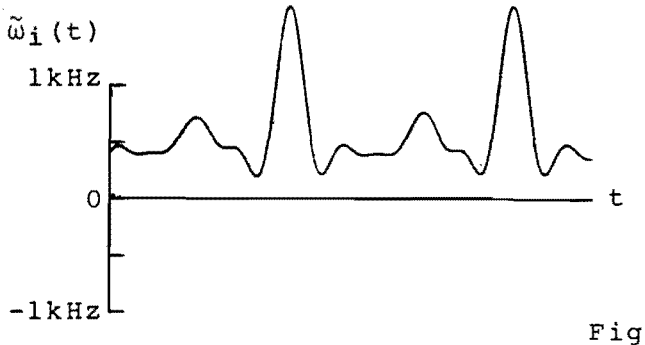


Fig 5.38.1

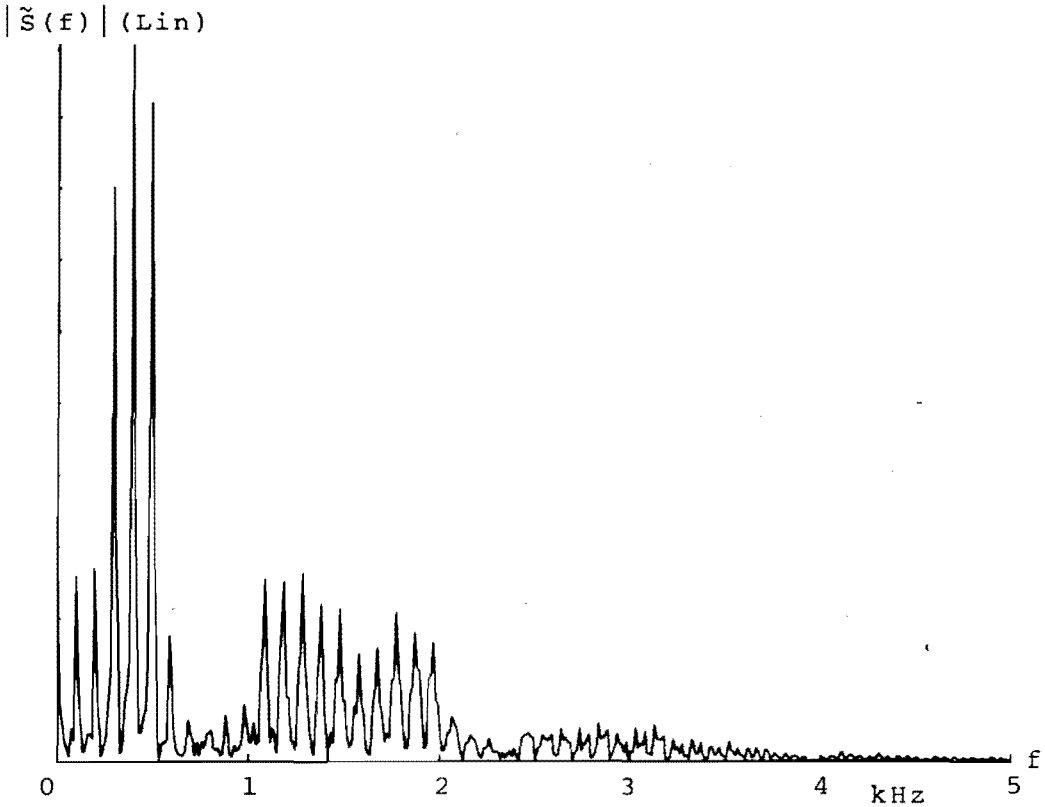


Fig 5.38.2

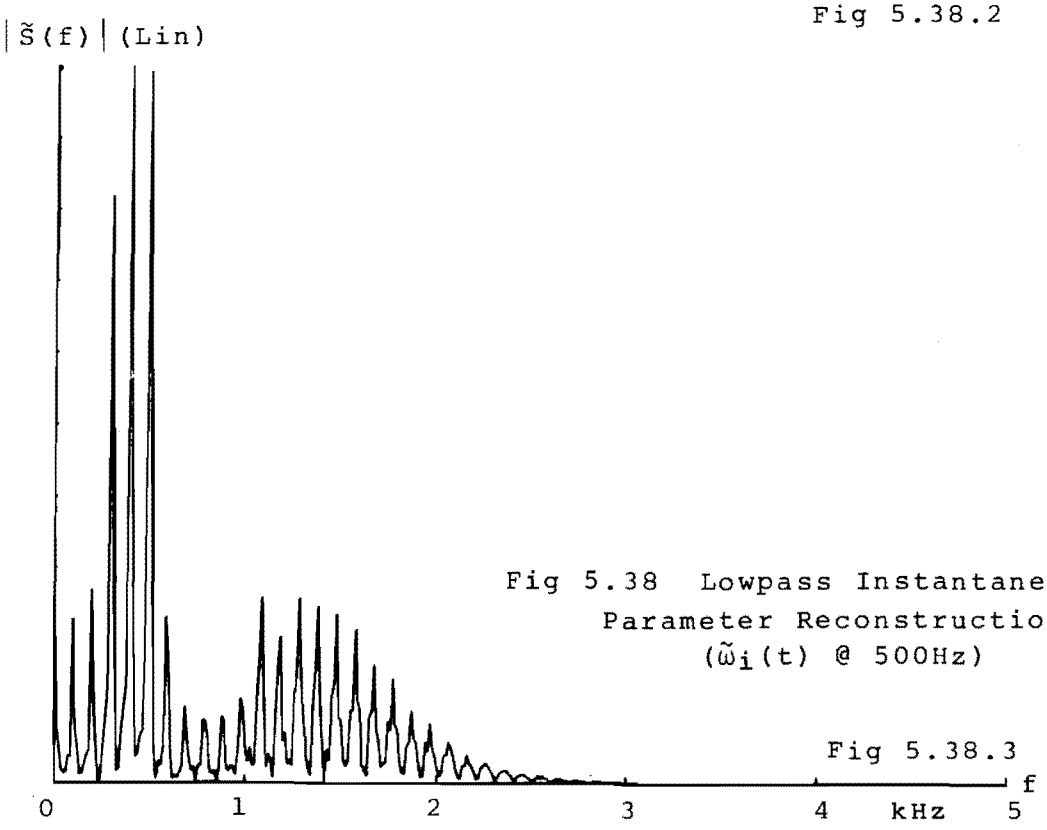


Fig 5.38.3

Fig 5.38 Lowpass Instantaneous Parameter Reconstructions ($\tilde{\omega}_i(t)$ @ 500Hz)

The fact that the waveform reconstructed from lowpass instantaneous amplitude and $\omega_i(t)$ lowpass filtered to 1,000 Hz appears to preserve the zero crossings of the original signal suggests that such a reconstruction may exhibit intelligibility similar to that of clipped speech. It should be noted, however, that slight modifications of zero crossing positions can seriously affect the intelligibility of this type of waveform (Ref. 119).

(5.7.2) UNVOICED FRICATIVE

Figure (5.39) is the amplitude spectrum of a band-limited version of the unvoiced fricative /ʃ/ uttered by a male speaker. It has been shown, Section (4.3.3.2), that the spectral characteristics of such a signal are almost fully described by the statistics of its instantaneous frequency function, and this suggests that constant amplitude reconstruction should cause little distortion. Figure (5.40) is the amplitude spectrum of a constant amplitude reconstruction of /ʃ/ and comparison with figure (5.39) reveals a slight bandwidth increase and some symmetricalisation of the spectrum around its centre frequency, ω_m.

An approximate constant frequency reconstruction of the phoneme with full bandwidth instantaneous amplitude and $\omega_i(t)$ lowpass filtered to 500 Hz generates the amplitude spectrum, figure (5.41). Although the magnitudes of components surrounding the central peak have been reduced and the spectrum made more symmetrical, this is still a good approximation of the original phoneme.

Lowpass filtering the instantaneous frequency functions of an unvoiced fricative will alter the statistics and shape of the associated pdf. Figures (5.42.1) to (5.42.5) are the pdfs of the full bandwidth instantaneous frequency waveform

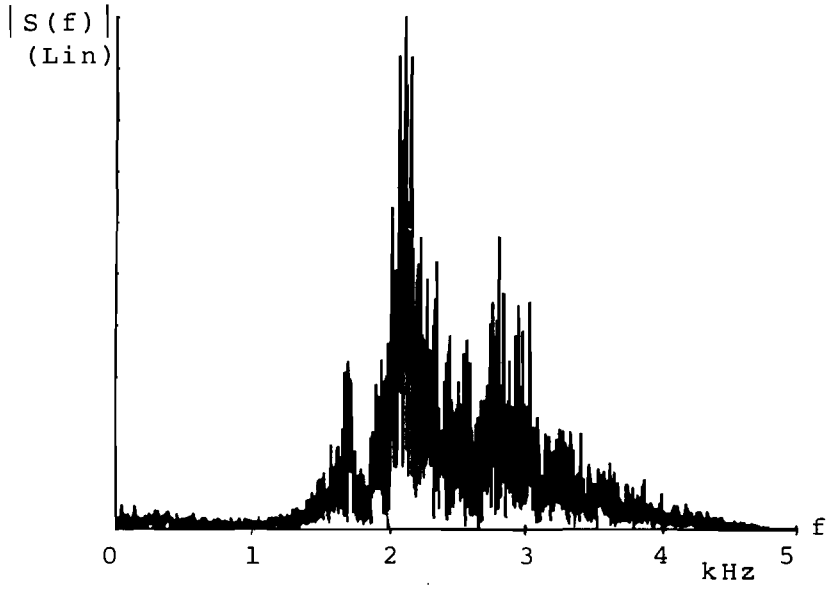


Fig 5.39 Unvoiced Fricative /ʃ/

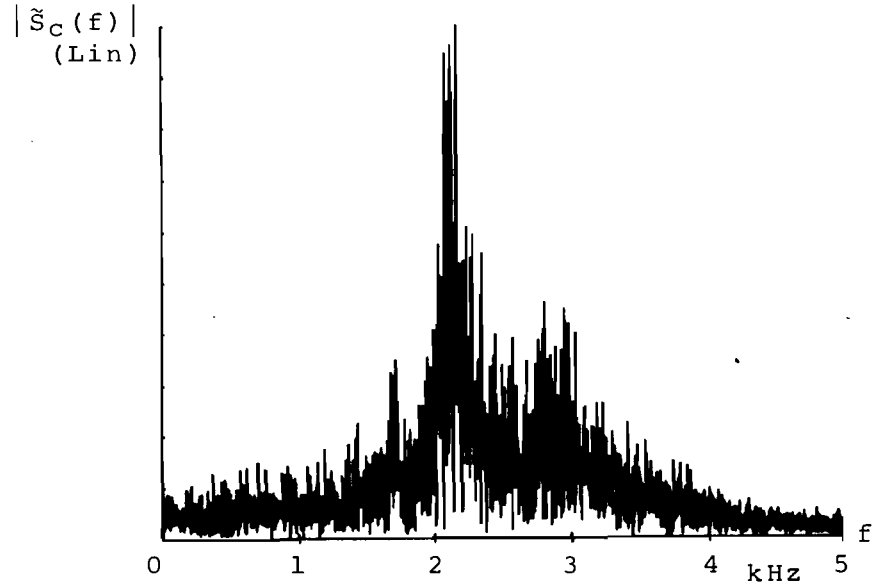


Fig 5.40 Constant Amplitude Reconstruction of /ʃ/

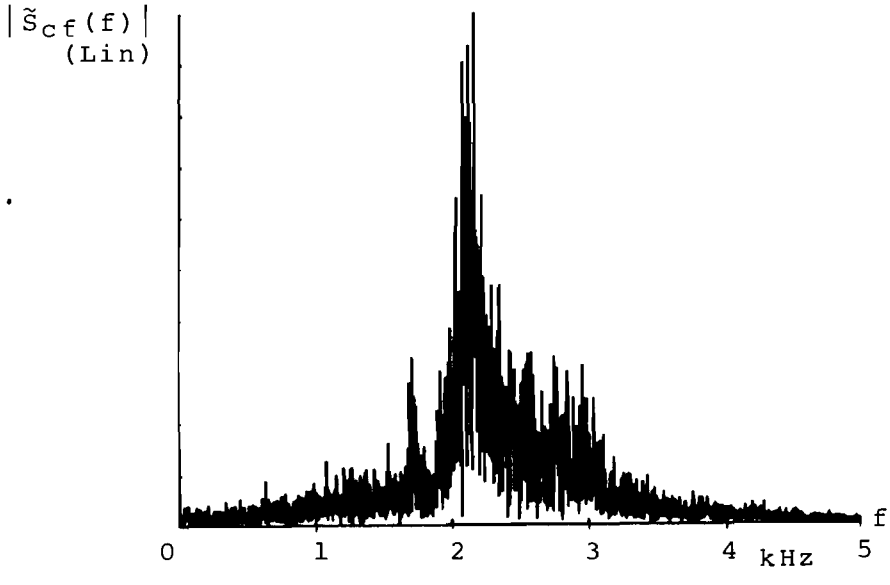


Fig 5.41 Full Bandwidth a(t), $\omega_i(t)$ @ 500Hz Reconstruction

of /ʃ/, and instantaneous frequency lowpass filtered to 2,000 Hz, 1,000 Hz, 500 Hz and 100 Hz respectively. The general trend is for a narrowing of the pdf around the centre frequency as the cutoff frequency is reduced, but the pdfs figures (5.42.3) and (5.42.4) exhibit definite assymetry. The "hump" in these pdfs at approximately 1 kHz above the centre frequency is reminiscent of the spectral envelope of the original phoneme, figure (5.39).

Narrowing of the reconstruction signal amplitude spectrum similar to that predicted by the narrowing of associated instantaneous frequency pdfs is illustrated by the set of amplitude spectra, figure (5.43). Figure (5.43.1) is the amplitude spectrum resulting from signal reconstruction with both a(t) and $\omega_i(t)$ lowpass filtered to 2,000 Hz, and this is a good approximation of the original.

Reconstructing with both a(t) and $\omega_i(t)$ lowpassed to 1,000 Hz generates figure (5.43.2) which exhibits a definite reduction in the magnitude of components surrounding the central peak. Figure (5.43.3) is the amplitude spectrum produced by reconstruction with a(t) and $\omega_i(t)$ lowpassed to 500 Hz, and there is now a definite reduction in signal bandwidth.

The effect of lowpass filtering the instantaneous frequency waveform on the pdf of the unvoiced fricative gives a good indication of the effects of such filtering on recontruction signal bandwidth. Such reduction of the phonemes bandwidth around its centre frequency will cause it to sound more like a "whistle" at frequency ω_m. The effects on intelligibility, however, will be dependent on the context of the phonemes use.

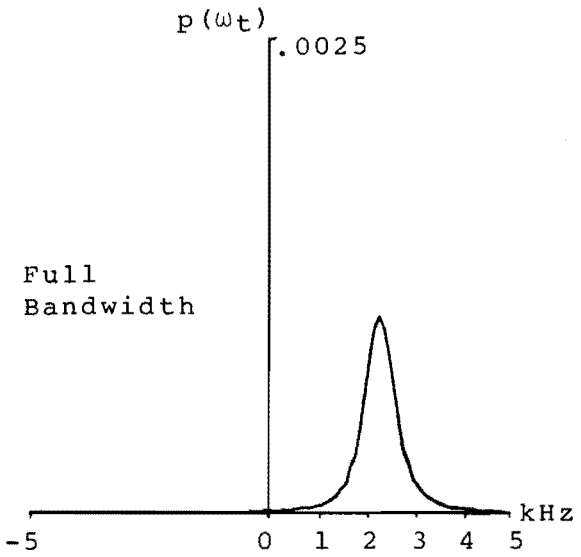


Fig 5.42.1

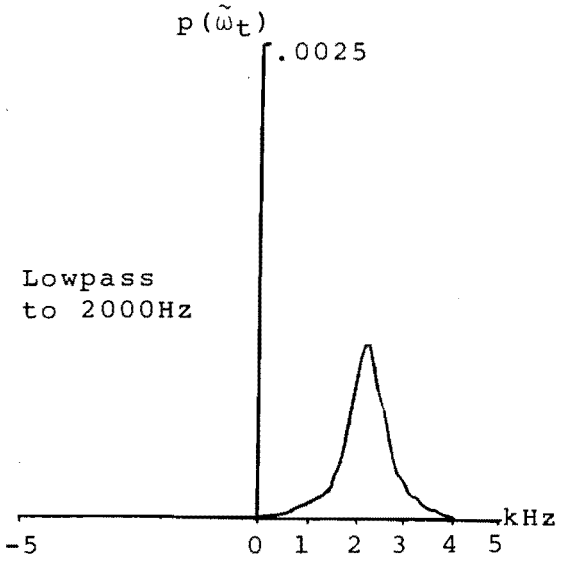


Fig 5.42.2

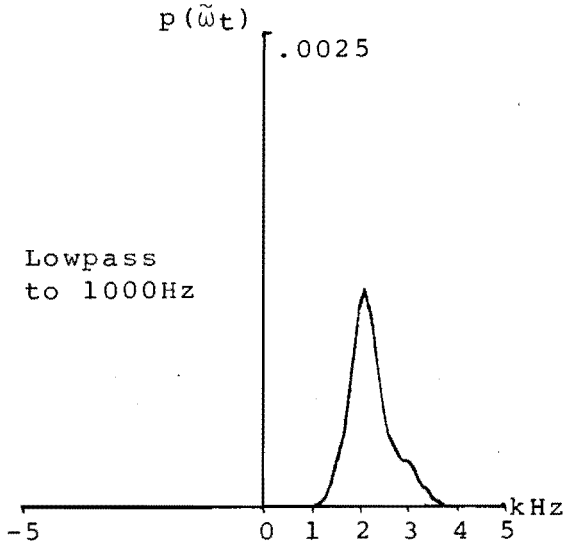


Fig 5.42.3

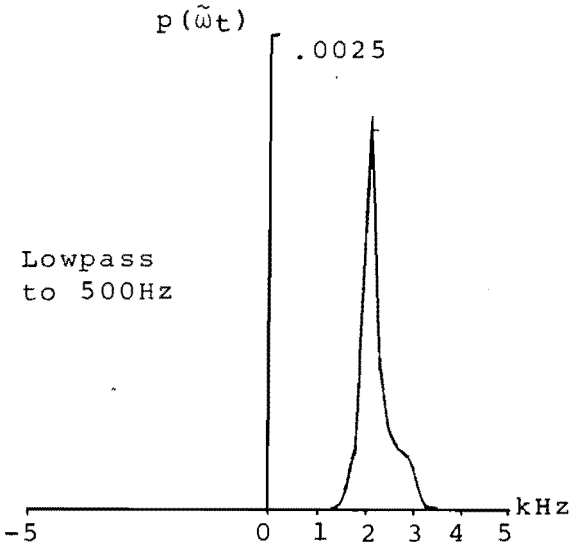


Fig 5.42.4

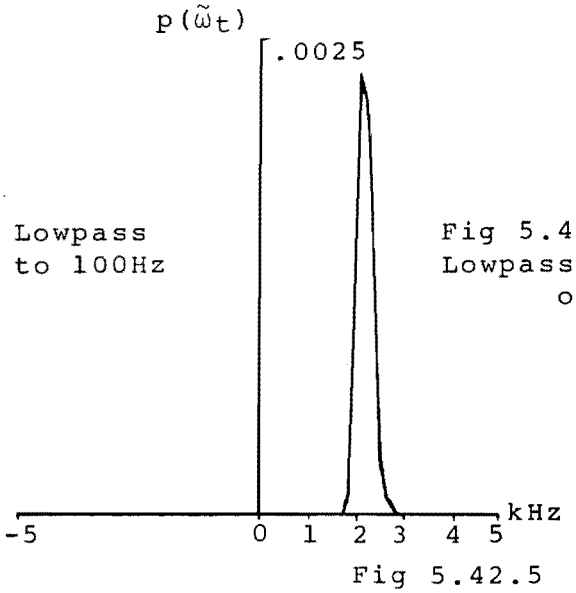


Fig 5.42.5

Fig 5.42 Effect of Lowpass Filtering $\omega_i(t)$ on its pdf

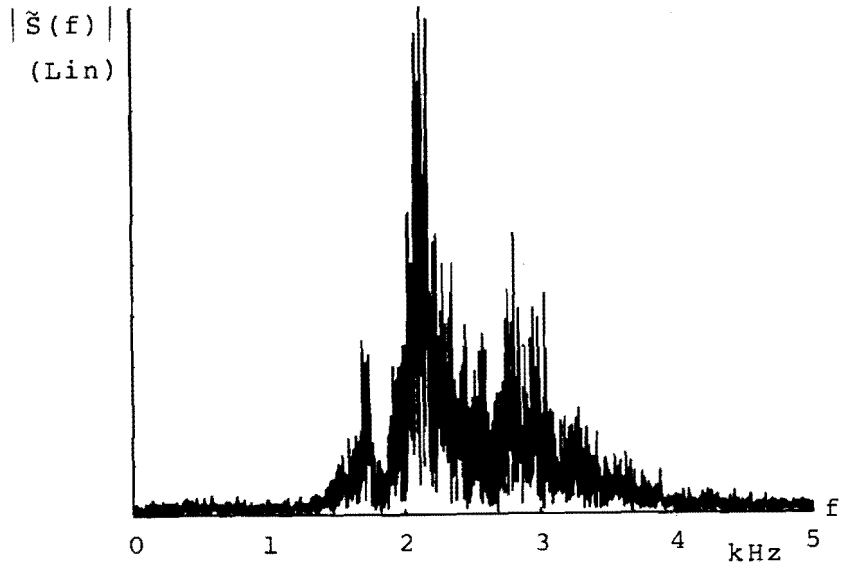


Fig 5.43.1 Reconstruction, a(t) and $\omega_i(t)$ Lowpass to 2000Hz

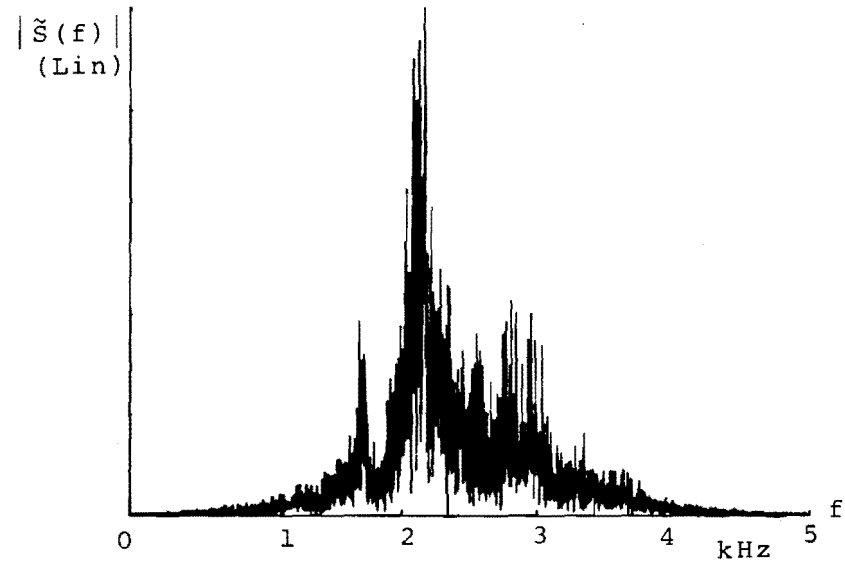


Fig 5.43.2 Reconstruction, a(t) and $\omega_i(t)$ Lowpass to 1000Hz

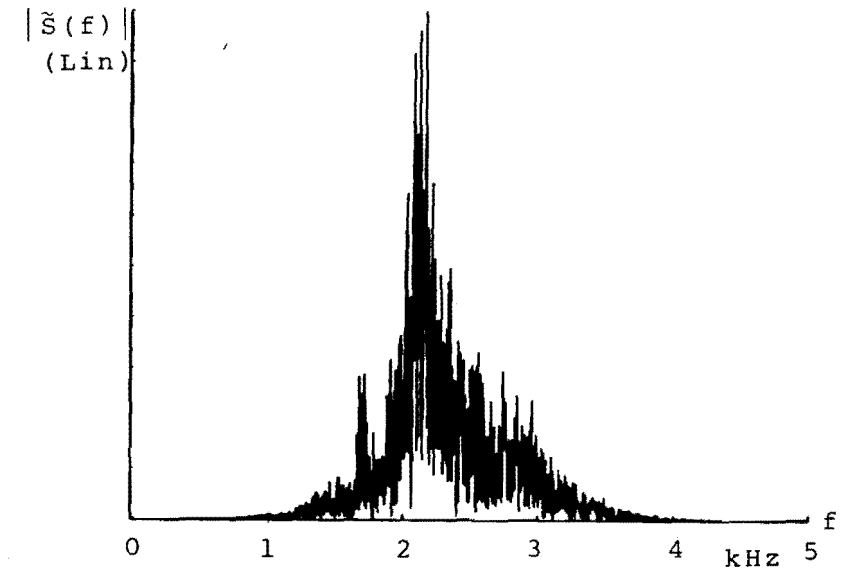


Fig 5.43.3 Reconstruction, a(t) and $\omega_i(t)$ Lowpass to 500Hz

(5.7.3) VOICED FRICATIVE

Figure (5.44) displays the logarithmic amplitude spectrum and 3 cycles of the time waveform and instantaneous functions of the voiced phoneme /z/. The instantaneous parameters of this phoneme have been studied in detail in Section (4.3.3), and it is known that the amplitude of the noise component of the phoneme changes from cycle to cycle.

The relatively high amplitude noise component of cycle 3 results in the generation of several UHP analytic signal zeros during the period of low instantaneous amplitude. This causes the pdf of instantaneous frequency to resemble that of an unvoiced phoneme over this low amplitude period and results in a short term change of the average instantaneous frequency.

Lowpass filtering both instantaneous frequency and amplitude to 1,000 Hz, figures (5.45.1) and (5.45.2), preserves the average instantaneous frequency change, and reconstruction from these lowpass parameters leads to the time waveform and time averaged amplitude spectrum, figures (5.45.3) and (5.45.4).

Comparison of the reconstruction and original waveforms, figures (5.45.3) and (5.44.2) reveals that only the high frequency waveform fluctuations which generated UHP analytic signal zeros have been preserved. (i.e. fluctuations which cause a pair of zero crossing for both $s(t)$ and $\hat{s}(t)$ are preserved). The number of such fluctuations per cycle is dependent on the relative amplitude of the phonemes noise component and slight variations of this amplitude have caused gross aperiodicity in the reconstruction waveform.

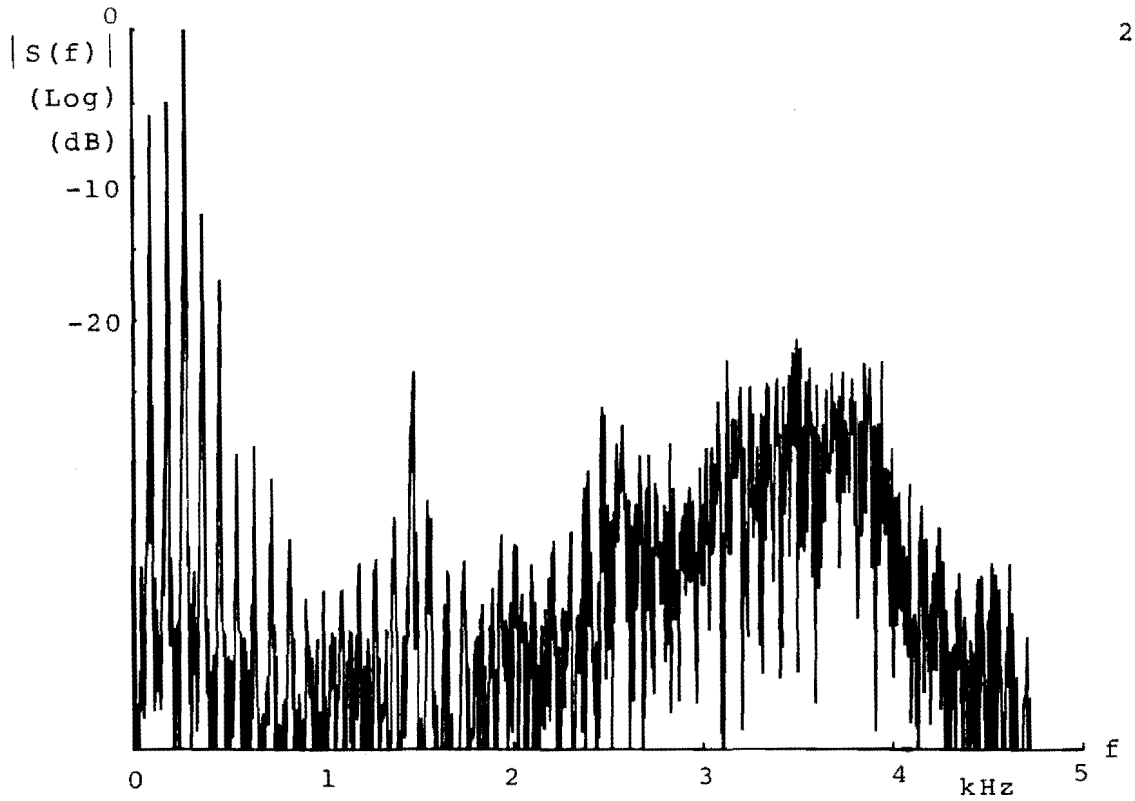


Fig 5.44.1

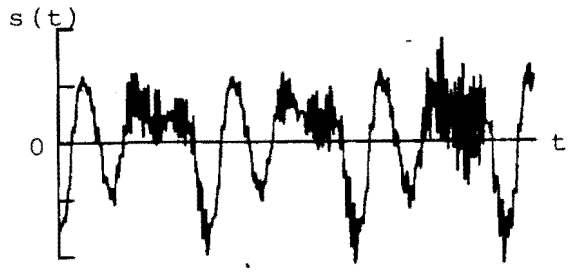


Fig 5.44.2

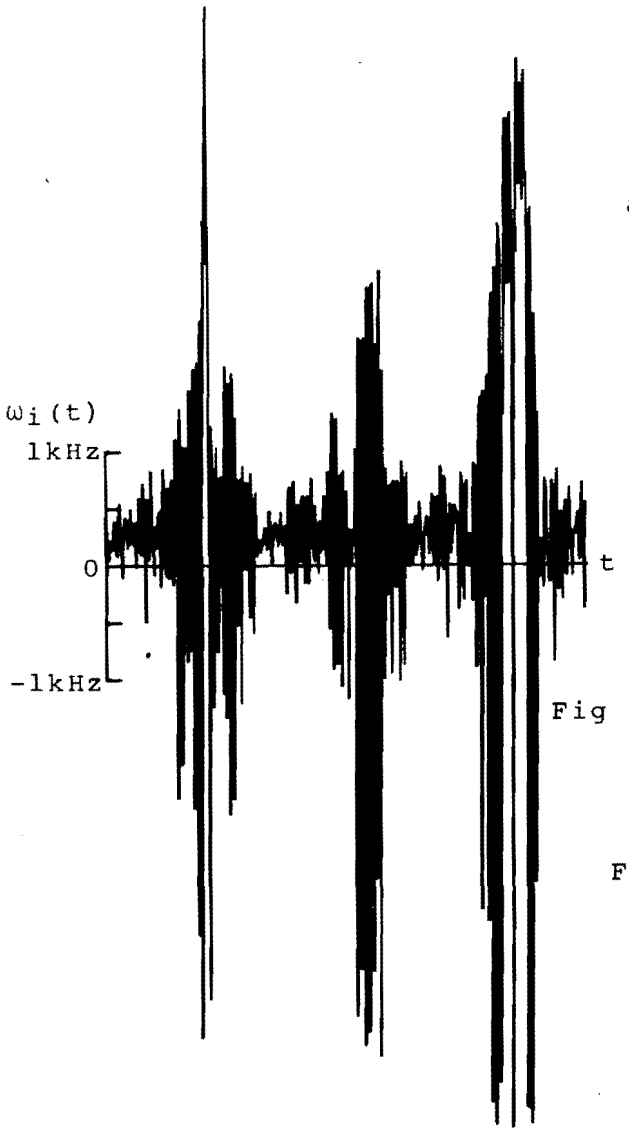


Fig 5.44.3

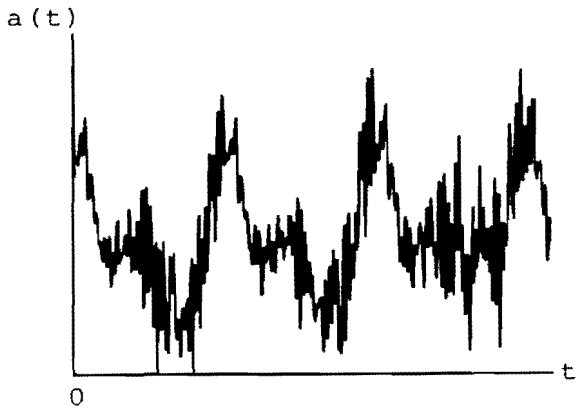


Fig 5.44.4

Fig 5.44 Voiced Fricative /z/

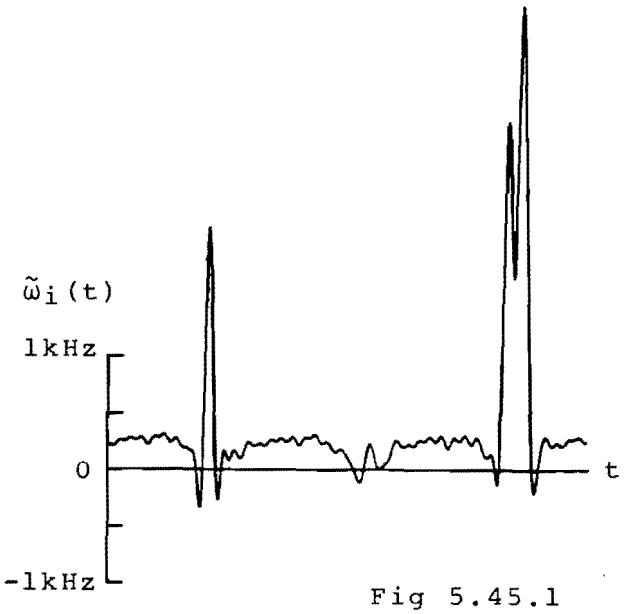


Fig 5.45.1

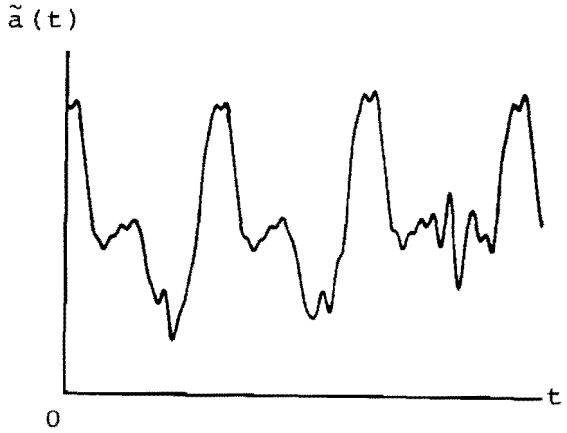


Fig 5.45.2

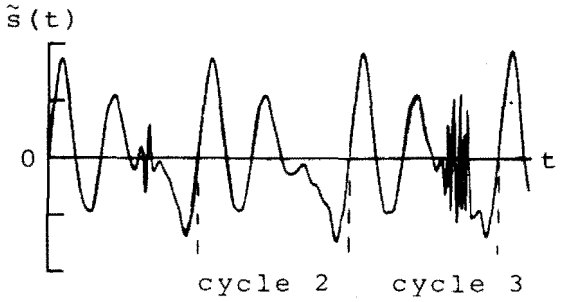


Fig 5.45.3

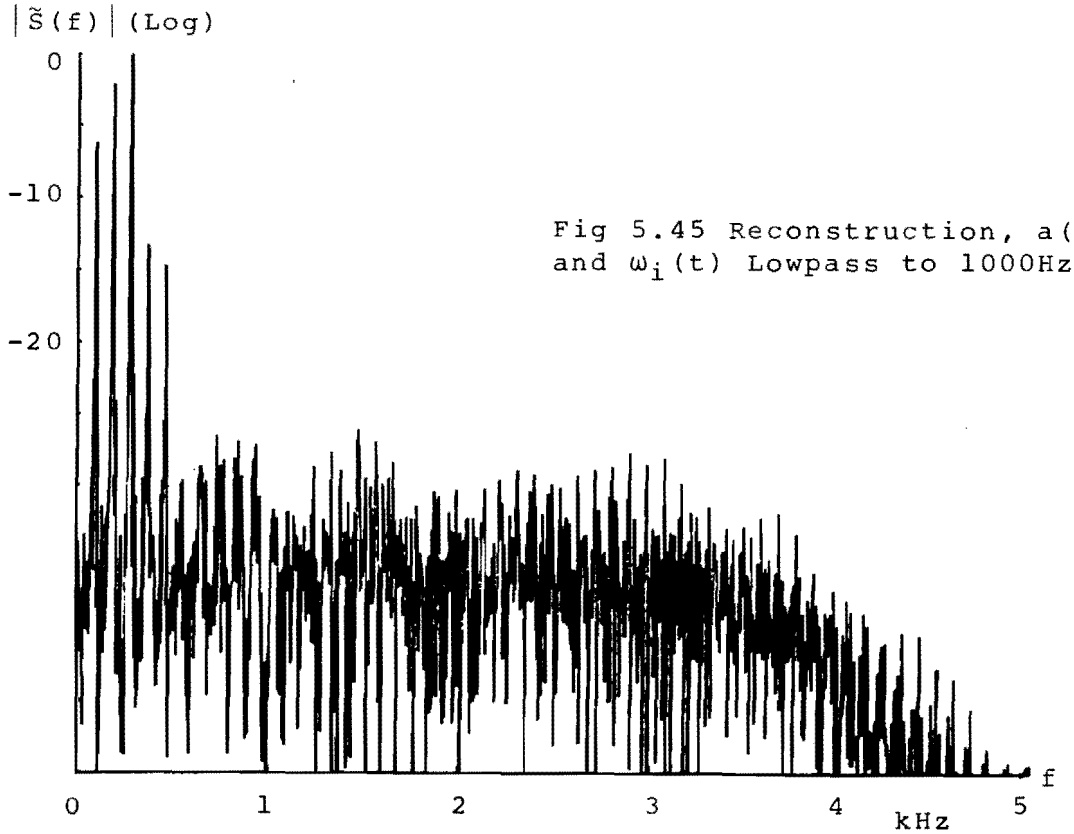


Fig 5.45 Reconstruction, a(t) and $\omega_i(t)$ Lowpass to 1000Hz

Fig 5.45.4

The reconstruction waveform indicates that the amplitude spectrum of the reconstruction will vary considerably from cycle to cycle. This can be studied by forming periodic waveforms from cycles 2 and 3 of figure (5.45.3).

Figures (5.46.1) and (5.46.2) are the periodic waveform and amplitude spectrum formed from "cycle 2". The spectrum is a good reconstruction of the low frequency spectral lines of the original amplitude spectrum, figure (5.44.1).

Forming a periodic waveform from "cycle 3" results in the waveform and amplitude spectrum, figures (5.47.1) and (5.47.2). This spectrum is also a reasonably good reconstruction of the low frequency spectral lines, but displays a great deal of high frequency energy which was not present in "cycle 2".

Although the time averaged amplitude spectrum of the reconstruction, figure (5.45.4), appears to be a good approximation to the original, this significant amplitude modulation of high frequencies from cycle to cycle must introduce a form of distortion.

(5.7.4) TEST PHRASE (1)

To further illustrate the distortion of particular phonemes and to investigate the effect on overall speech intelligibility, the phase "fast zoo" has been subjected to various lowpass instantaneous parameter reconstructions. These are presented in Part 2 of the cassette tape which accompanies this thesis.

Version (a) is the original phrase lowpass filtered to 3,400 Hz. This forms the reference data with which subsequent versions should be compared.

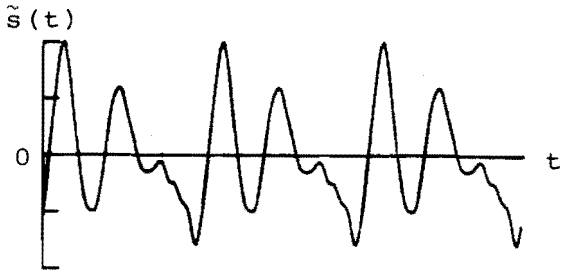


Fig 5.46.1

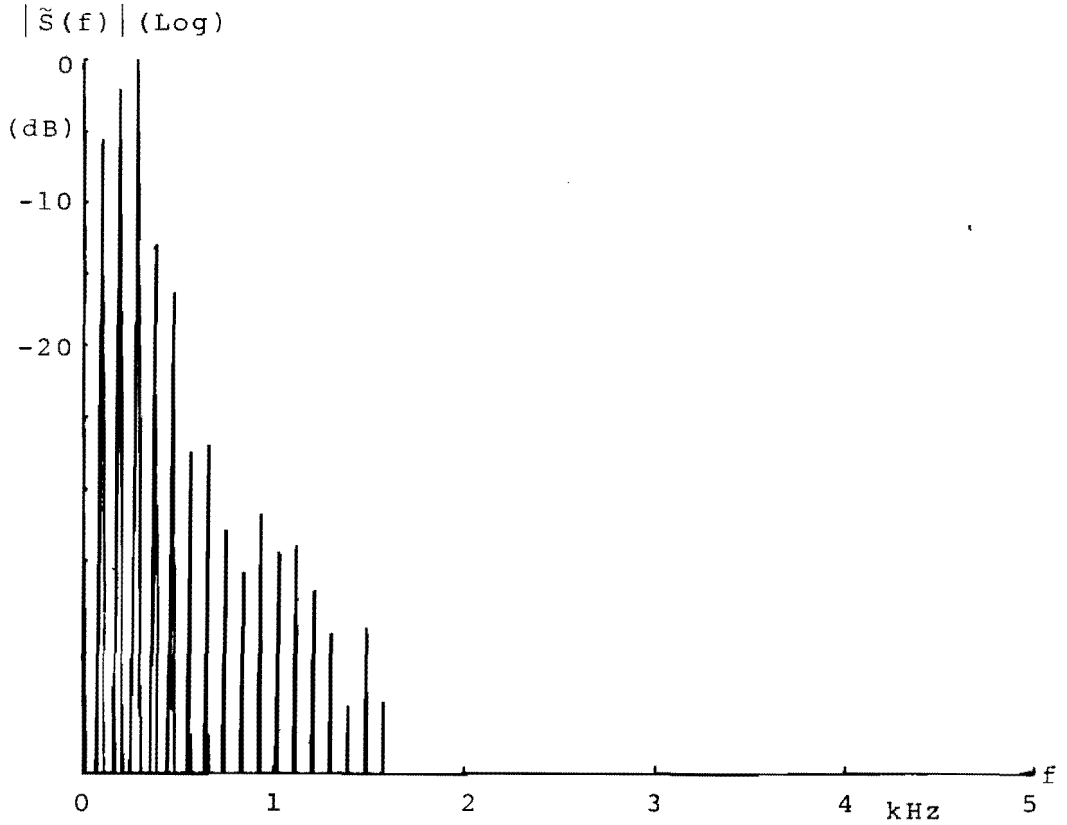


Fig 5.46.2

Fig 5.46 Analysis of "cycle 2"

Fig 4.47.1

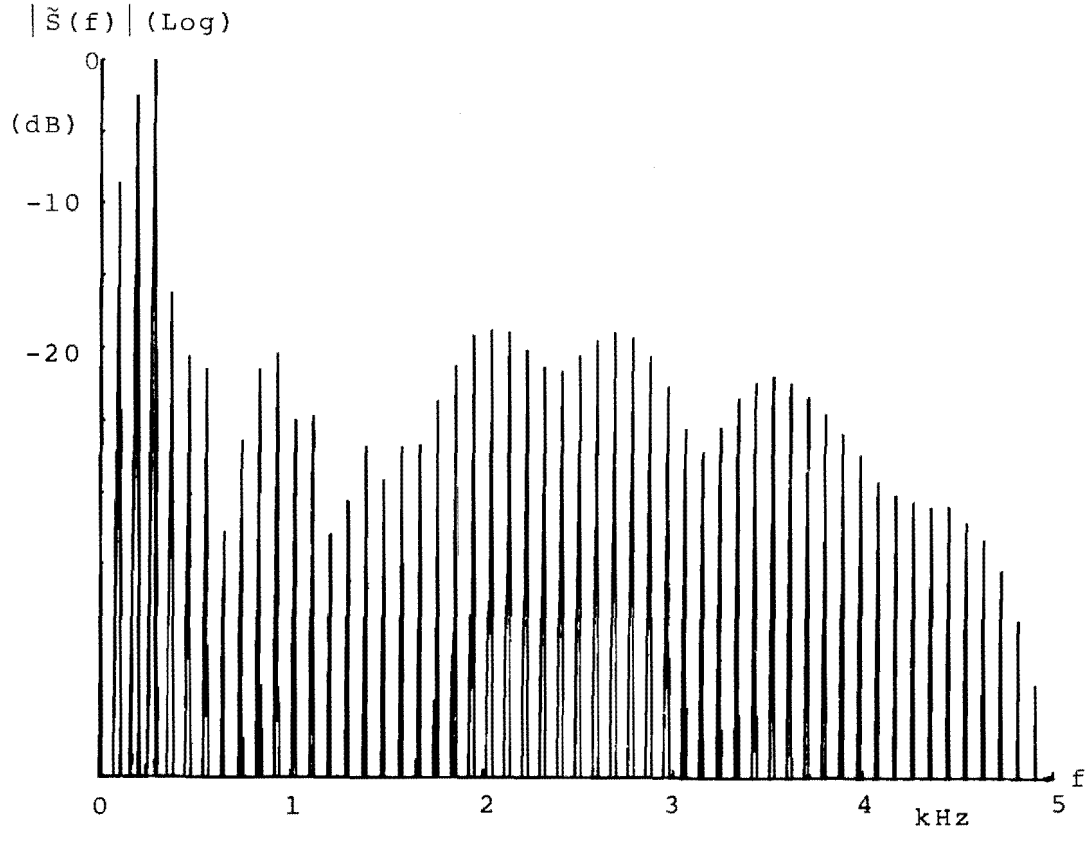


Fig 5.47.2

Fig 5.47 Analysis of "cycle 3"

In order to demonstrate its relative intelligibility, version (b) is an approximation to constant amplitude speech, reconstructed from full bandwidth $\omega_i(t)$ and a(t) lowpass filtered to 500 Hz. Conversely, version (c) is an approximation to constant frequency speech and is reconstructed from full bandwidth a(t) and $\omega_i(t)$ lowpass filtered to 500 Hz. Although version (c) is intelligible, it exhibits a form of "burbling" distortion which seems to be present during most phonemes. Initial listening tests have shown versions (b) and (c) to be more intelligible when replayed through free field speakers rather than headphones.

Version (d) is the reconstruction from both a(t) and $\omega_i(t)$ lowpass filtered to 1,000 Hz. Once again, most phonemes exhibit a form of dynamic distortion which is characterised by "burbling" of unvoiced fricatives and vowels, and the voiced phoneme /z/ is "broken" by random bursts of high frequency noise. Version (e), created from a(t) and $\omega_i(t)$ lowpass filtered to 500 Hz displays an even greater degree of "burbling".

The fact that reconstructed phonemes, other than /z/, exhibit forms of dynamic distortion suggests that they also possess rapidly fluctuating values of average instantaneous frequency. Such fluctuations of $\overline{\omega_i(t)}$ have been demonstrated for the quasi-stationary vowel sound /ɛ/, Section (4.3.1.2), and the effects of such changes on a lowpass instantaneous parameter vowel reconstruction are now studied in detail.

(5.7.5) APERIODIC VOWEL ANALYSIS

Figure (5.48.1) is the time averaged amplitude spectrum of the bandpass vowel /ɛ/ uttered by a male speaker. Blurring of the spectral line structure at high frequency indicates that the vowel is not perfectly periodic, and that it may exhibit slight changes of pitch and of the second and

third formant amplitudes. Three cycles of the vowels instantaneous parameters are illustrated in figures (5.48.2) and (5.48.3).

As both real and orthogonal time signals, figures (5.48.4) and (5.48.5) exhibit different zero crossing counts for each of the three cycles displayed, average instantaneous frequency can be expected to change from cycle to cycle. This is confirmed by lowpass filtering the instantaneous parameters to 1,000 Hz, figures (5.48.6) and (5.48.7).

Reconstruction from these lowpass parameters yields the time waveform segment, figure (5.48.8), which is grossly aperiodic. Analysis of long segments of the reconstruction time waveform led to the discovery that it consisted principally of three basic "cycles". A representative of each of these has been used to form a "periodic" version of the reconstruction and the results are presented as figures (5.49) to (5.51).

Fourier analysis of the waveform, figure (5.49.1), shows it to be a good approximation of the first formant of the original vowel. The reconstruction "cycle" used to form this signal exhibited the lowest average instantaneous frequency of the three analysed.

The waveform, figure (5.50.1), exhibits considerable spectral energy at each of the first, second and third formant positions, but its spectral envelope is not a good copy of the original.

Figure (5.51.1) is formed from the "cycle" which exhibited the highest average instantaneous frequency of the three (corresponding to that of the 10th harmonic by zero crossing count). In this case, the reconstruction amplitude spectrum is a very good approximation of the original, but with higher than average second and third

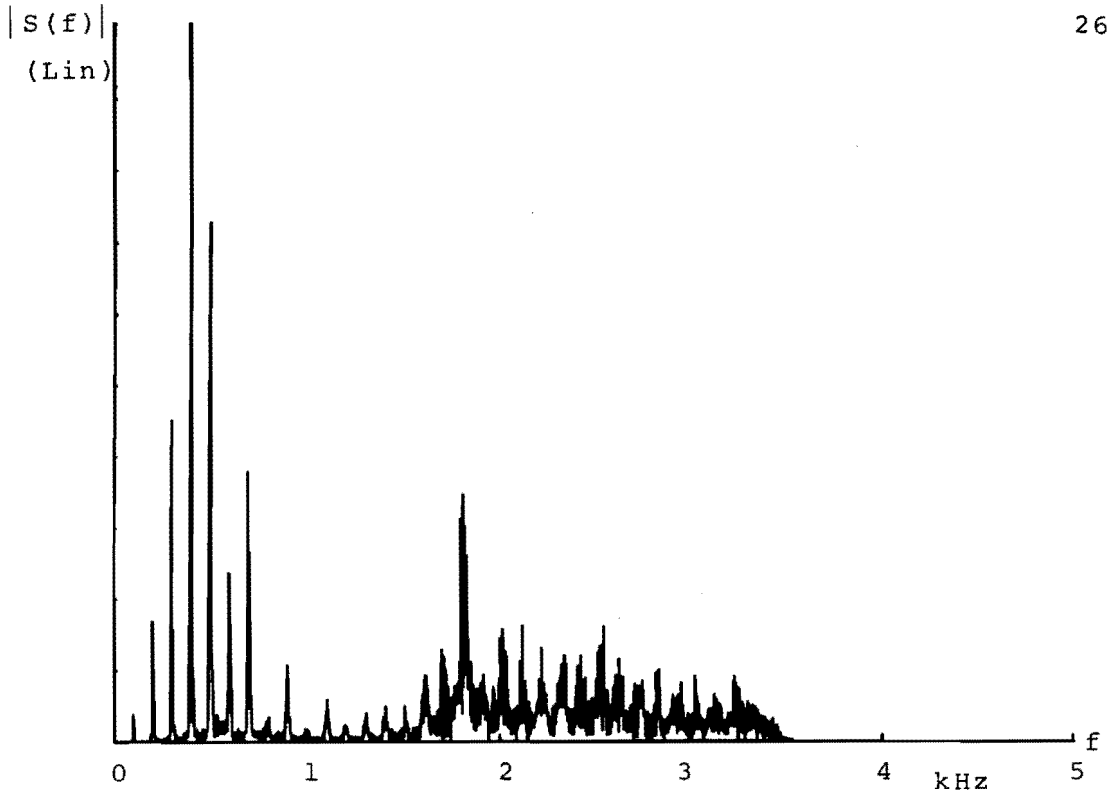


Fig 5.48.1

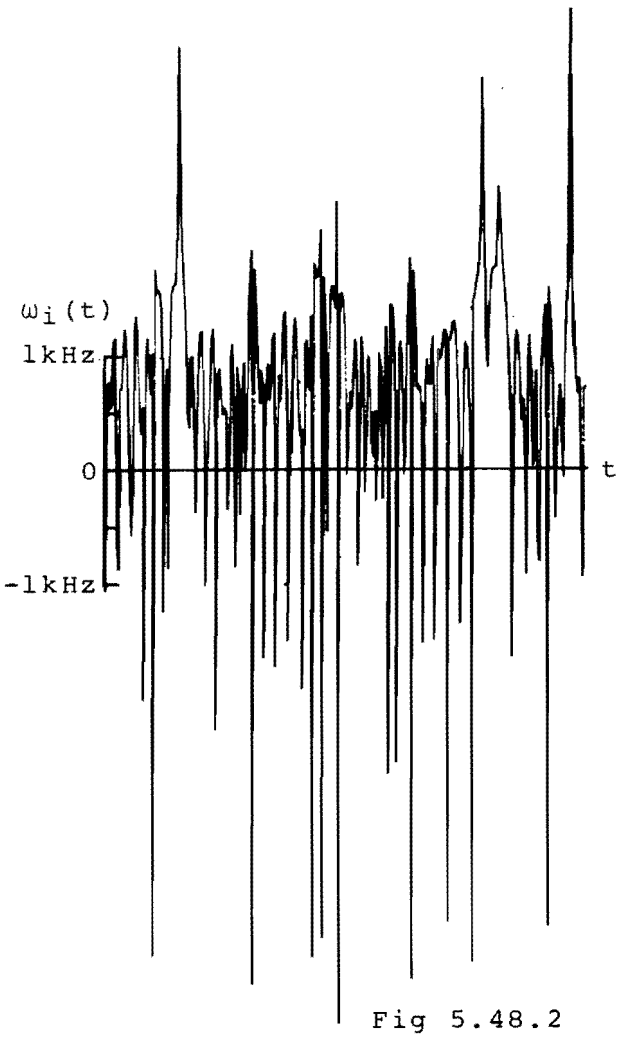


Fig 5.48.2

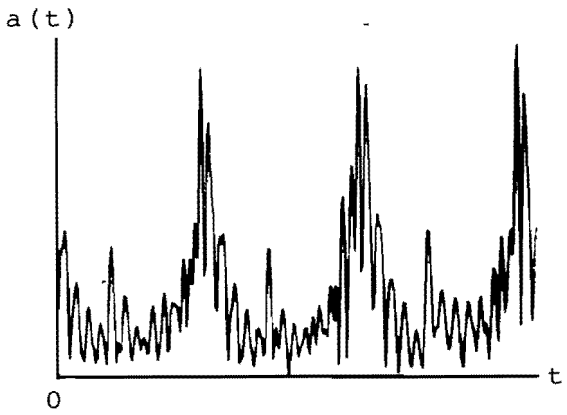


Fig 5.48.3

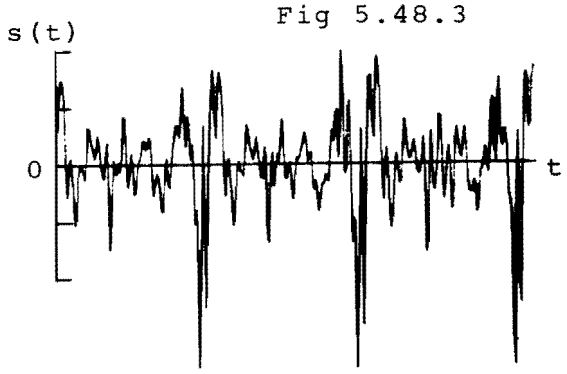


Fig 5.48.4

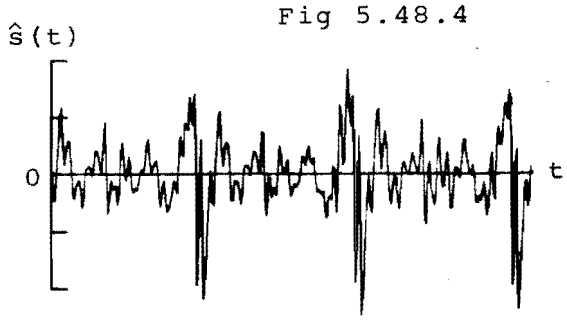


Fig 5.48.5

Figs 5.48.1-5.48.5 Analysis of Aperiodic Vowel

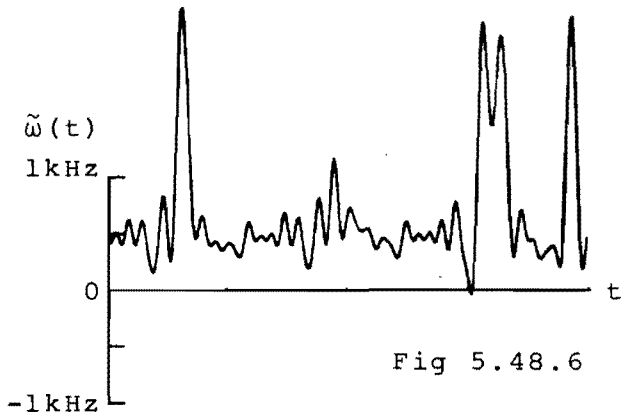


Fig 5.48.6

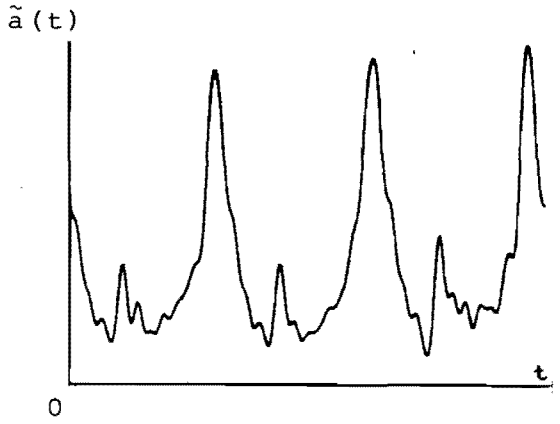


Fig 5.48.7

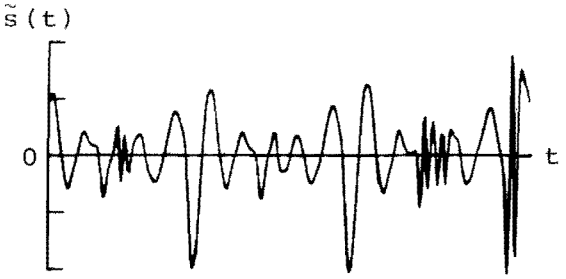


Fig 5.48.8

Figs 5.48.6-5.48.8 Reconstruction from Instantaneous Parameters Lowpass Filtered to 1000Hz

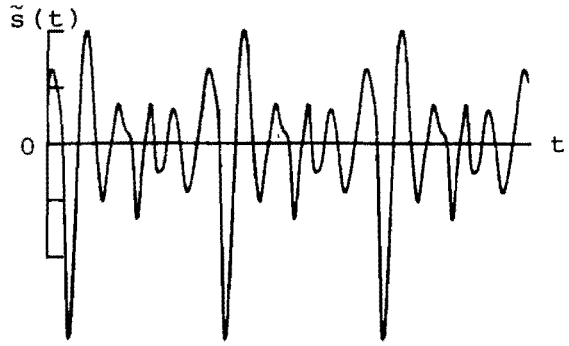


Fig 5.49.1

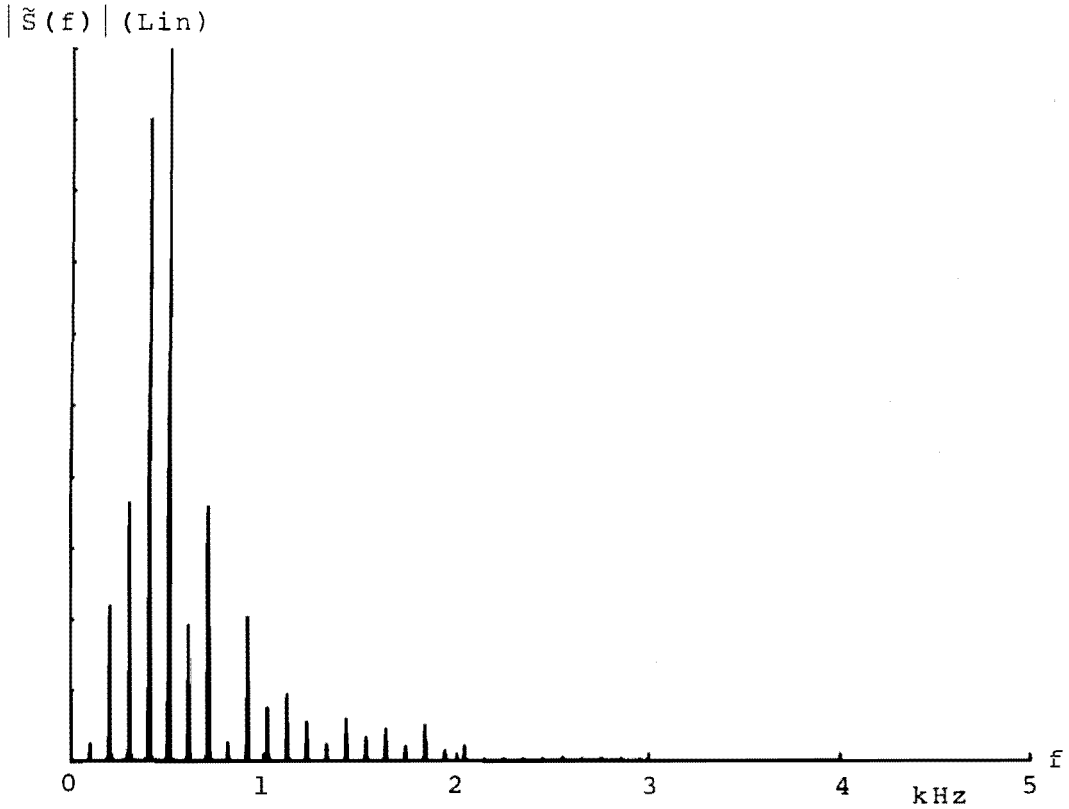


Fig 5.49.2

Fig 5.49 Analysis of "Cycle A"

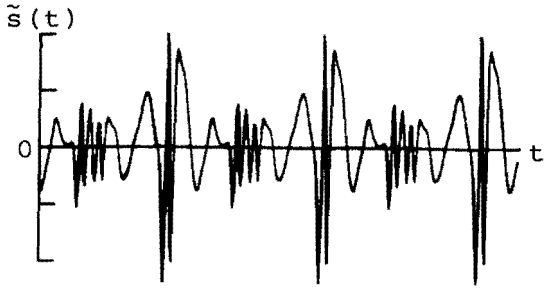


Fig 5.50.1

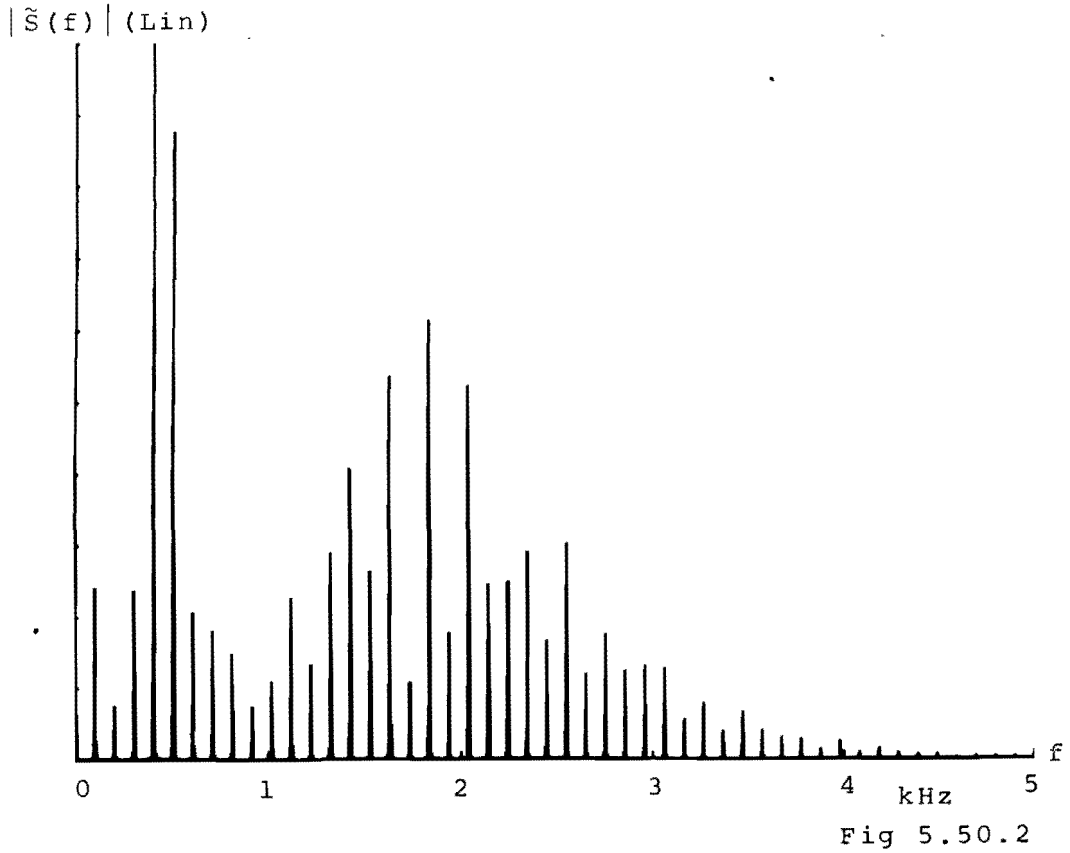


Fig 5.50.2

Fig 5.50 Analysis of "Cycle B"

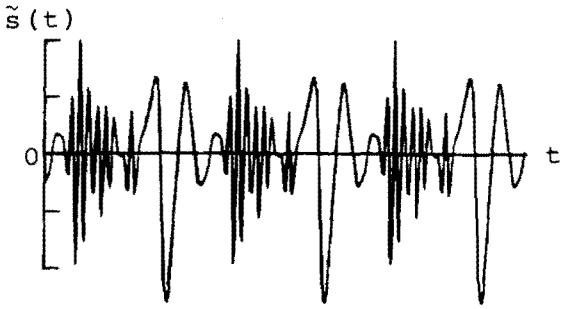


Fig 5.51.1

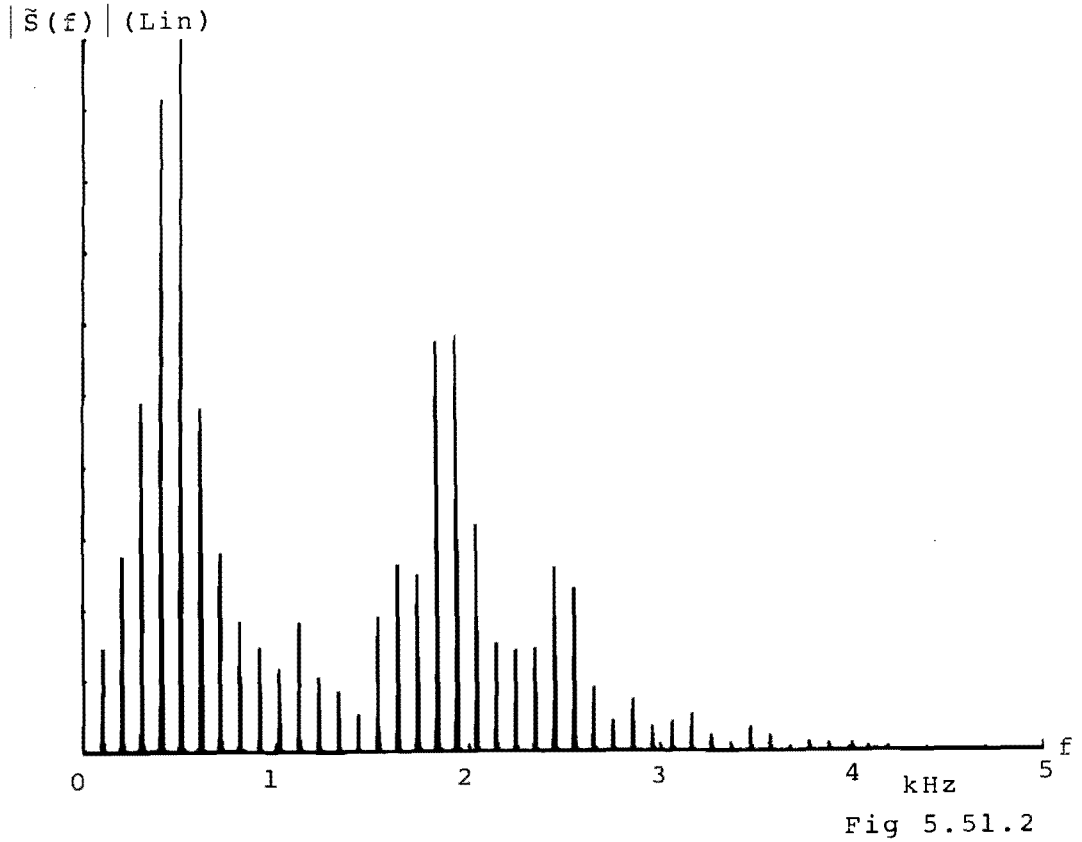


Fig 5.51.2

Fig 5.51 Analysis of "Cycle C"

formant amplitudes. Analysis of the instantaneous parameters of a differentiated vowel, Section (5.4.1), confirms that increasing the amplitudes of second and third formants causes average instantaneous frequency to rise.

The results of the above analysis have two major consequences. The first is that, unless perfectly periodic, lowpass instantaneous parameter reconstruction of a vowel will result in a dynamic form of distortion which involves amplitude modulation of the reconstructions high frequency components. Part 3 of the cassette tape containes a version of the original vowel of figure (5.48) followed by a reconstruction with the instantaneous parameters lowpass filtered to 1,000 Hz. Modulation of the reconstructions high frequency components is evident in the second version and this distortion resembles the "burbling" observed during lowpass instantaneous parameter reconstruction of the test phrase "fast zoo". It is apparent that any phoneme which exhibits random fluctuations of its average instantaneous frequency value will be dynamically distorted by lowpass instantaneous parameter reconstruction.

Accurate reconstruction of the vowel spectrum by the lowpass instantaneous parameter reconstruction waveform, figure (5.51.1), was not predicted as lowpass filtering of both instantaneous amplitude and frequency to 1,000 Hz should have removed fluctuations carrying bandwidth, and thus second and third formant, information. As noted previously, however, lowpass filtering $\omega_i(t)$ to 1,000 Hz maintains the short term averages of instantaneous frequency and, correspondingly, the resulting lowpass instantaneous parameter reconstruction will exhibit good approximations to all of the zero crossings of the original waveform which were due to UHP analytic signal zeros. Versions of vowels which exhibit high average instantaneous frequency also exhibit many zero

crossings due to UHP analytic signal zeros and these are all reproduced by this lowpass instantaneous parameter reconstruction.

This leads to the second consequence of the above results which is that versions of vowels which exhibit high average instantaneous frequency will lose less "information" in a lowpass instantaneous parameter reconstruction than vowels with a low value of average instantaneous frequency. For this reason, the intelligibility of a lowpass instantaneous parameter reconstruction of a differentiated vowel may be compared with the known "high intelligibility" of a clipped differentiated vowel. The two are not identical, however, as the clipped waveform exhibits some zero crossings which are due to analytic vector inner loops (LHP zeros causing negative instantaneous frequencies).

(5.7.6) TEST PHRASE (2)

The quality and intelligibility of lowpass instantaneous parameter reconstruction of differentiated speech is tested using the pre-emphasized speech segment, Version (a), from Part (4) of the accompanying cassette tape. Prior to pre-emphasis this speech segment was bandlimited to the "telephone bandwidth". Table 5.1 shows the parameter bandwidths used for each reconstruction.

VERSION	a(t) BANDWIDTH	$\omega_i(t)$ BANDWIDTH
b	1000 Hz	full
c	200 Hz	full
d	constant	full
e	full	1000 HZ
f	full	500 Hz
g	full	200 Hz
h	1000 Hz	1000 Hz
i	500 Hz	500 Hz
j	200 Hz	200 Hz
k	squelch	1000 Hz
l	squelch	500 Hz
m	squelch	100 Hz
n	200 Hz	1000 Hz

Table 5.1

No de-emphasis has been applied to the reconstructions before recording.

Versions (b) to (d) demonstrate the effect of instantaneous amplitude bandwidth reduction on reconstruction fidelity. Although there is noticeable degradation of speech quality, Versions (b) and (c) remain intelligible. The "constant amplitude" reconstruction Version (d), however, is severely degraded by the introduction of noise in the pauses between words and phonemes.

Versions (e) to (g) demonstrate the effect of successively more severe lowpass filtering of instantaneous frequency on reconstruction quality. The resulting

distortions are only just noticeable in Version (e) but have increased to a severe "burbling" by Version (g). The relative qualities of the reconstructions, Versions (c) and (g), confirm that the robustness of speech intelligibility under loss or distortion of the instantaneous amplitude function does not apply equally to loss or distortion of the instantaneous frequency function.

All of the reconstructions, Versions (h) to (j) exhibit some degree of the "burbling" distortion. This varies from slight in Version (h) to extreme in Version (j), and the apparent intelligibility of Version (j) is reduced accordingly. Although some burbling is present, the quality of reproduction of voiced phonemes in Version (h) suggests that pre-emphasised speech has been well reproduced by this reconstruction.

Comparison of the relative qualities of Versions (f) and (i) reveals that maintenance of full bandwidth instantaneous amplitude in the performance of lowpass instantaneous frequency reconstructions improves the reconstruction quality. This confirms the observations of Sections (5.7.1) and (5.7.2).

The fourth set of reconstructions, Versions (k) to (m), are constant amplitude, lowpass instantaneous frequency with a simple squelch applied to remove noise between words.

It has been noted, Section (5.7.5), that lowpass filtering the instantaneous frequency function of speech to 1000 Hz appears to maintain short term changes of $\overline{\omega_i(t)}$, but usually removes the possibility of negative instantaneous frequency excursions. A reconstruction without amplitude information will, therefore, reproduce a signal whose informational attributes are zero crossing positions which are good approximations to those zero crossing

positions of the original speech signal which were due to UHP analytic signal zeros. This similarity between constant amplitude lowpass instantaneous frequency reconstruction and clipped speech is confirmed by Version (k) which exhibits intelligibility which is estimated to be as high as that reported for clipped, differentiated speech (Ref. 83).

Constant amplitude reconstructions with instantaneous frequency lowpass filtered to below 1000 Hz, Versions (l) and (m), exhibit reduced intelligibility. This may be explained in terms of the lowpass instantaneous frequency function and classical FM theory (Section (5.6)) or by the "spreading" of information bearing reconstruction zero crossing positions. Extreme lowpass filtering of instantaneous frequency results in a reconstruction whose zero crossing locations are spread to resemble those of a sinusoid.

The final lowpass instantaneous parameter reconstruction, Version (n), demonstrates that, although not as perceptually significant as instantaneous frequency, the inclusion of a small instantaneous amplitude bandwidth improves reconstruction quality.

(5.8) DISCUSSION

This Chapter has documented investigations into the properties of real signals reconstructed from original or modified versions of their instantaneous parameters, $a(t)$ and $\omega_i(t)$. Alteration of these parameters has been shown to cause distortion and spreading of reconstruction spectra, with reconstructed speech intelligibility proving to be more dependent on the fidelity of $\omega_i(t)$ than that of $a(t)$. No attempt has been made to produce intelligibility scores

for each type of reconstruction, but, where appropriate, relative intelligibilities have been compared.

Bandwidth efficient speech transmission based on narrow bandwidth instantaneous parameter reconstruction has been shown to be theoretically possible, provided $\omega_i(t)$ is coded with sufficient accuracy. The limitations of such a scheme are defined by the distortions inherent in the reconstructions.

Distortions resulting from lowpass instantaneous frequency reconstruction of speech are of two kinds. These are

(1) Distortion of the speech amplitude spectrum, usually resulting in loss of high frequency information, and

(2) "burbling".

For reconstruction with a given lowpass instantaneous frequency bandwidth, the first (static) form of distortion may be reduced by increasing the bandwidth of the lowpass instantaneous amplitude function used in reconstruction, or by pre-emphasis of the original speech signal prior to instantaneous parameter analysis.

The second (dynamic) form of distortion can be avoided only by speech pre-processing to prevent $\overline{\omega_i(t)}$ from changing during cycles of a voiced sound. Lowpass instantaneous frequency reconstructions of artificially periodic vowels, therefore, exhibit the first type of distortion, but no "burbling".

The intelligibility of speech reproduced from lowpass instantaneous parameters is most easily explained in terms

of the zeros of the associated analytic signal, $\Psi(z)$. Both $a(t)$ and $\omega_i(t)$ carry zero location information, but it is "encoded" differently in each case.

The zeros of the square of the instantaneous amplitude function $a^2(z)$, are the non-removed zeros of $\Psi(z)$ and $\Psi^*(z)$, and they may appear in the complex time plane as indicated in figure (5.52).

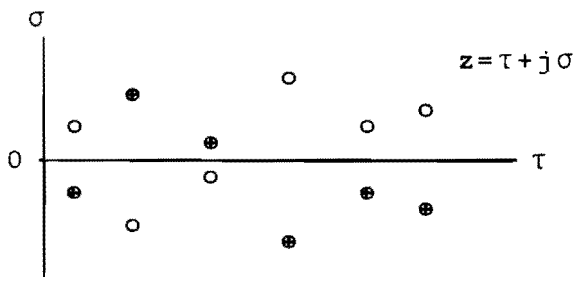


Figure (5.52) Zeros of $a^2(z)$

If the zeros of $\Psi(z)$ are those marked by "+", then it can be seen that recovery of $\Psi(z)$ from $a^2(z)$ requires choosing the correct zero from each conjugate pair. This problem is only trivial in the cases where $\Psi(z)$ is minimum or maximum phase and all of its zeros are LHP or UHP respectively.

Lowpass filtering $a^2(t)$ removes, or reduces the amplitude of, waveform fluctuations and is therefore equivalent to shifting zeros away from the real time axis. The zeros of lowpass $\tilde{a}^2(t)$ now correspond to $\tilde{\Psi}(z)$ and $\tilde{\Psi}^*(z)$ which are reduced bandwidth versions of $\Psi(z)$ and $\Psi^*(z)$ respectively.

Examination of the amplitude spectrum of a(t) generated by a vowel, figure (4.4.5), reveals that lowpass filtering removes information which corresponds to high frequency spectral components of the original vowel. If, as in the case of an unvoiced phoneme, average instantaneous frequency is nearer the centre of the analysis bandwidth, it may be treated as a carrier and lowpass filtering of a(t) results in loss of information symmetrically about $\overline{\omega_i(t)}$.

The instantaneous frequency function of speech codes the locations of the non-removed zeros of $\Psi(z)$ as fluctuations whose real time location, magnitude and polarity reflect τ, $|\sigma|$ and sgn (σ) respectively. The time average of instantaneous frequency is set by the combined rate of occurrence of removed zeros and non-removed UHP zeros.

Lowpass filtering of the instantaneous frequency waveform reduces the magnitude of some fluctuations. This may be viewed as shifting the locations of corresponding zeros (of $\Psi(z)$) away from the real time axis (increasing $|\sigma|$), thereby creating the lowpass waveform $\tilde{\omega}_i(t)$ whose fluctuations correspond approximately to the zeros of $\tilde{\Psi}(z)$, which is again a reduced bandwidth version of $\Psi(z)$.

Unfortunately, lowpass filtering of $\omega_i(t)$ not only causes a reduction in the magnitude of fluctuations, but the shapes of resulting lowpass fluctuations may be termed "distorted" as they are no longer members of the family of curves depicted by figures (2.4) and (2.5). The instantaneous frequency function of speech often exhibits many impulse like fluctuations, and convolution of $\omega_i(t)$ with an impulse response, h(t), results in a lowpass waveform which appears to be made up of scaled versions of h(t). This is illustrated for a vowel in figure (5.30).

A simple example of the effect that fluctuation shape distortion may have on the resulting reconstruction spectrum are the in-band spectral components caused by the periodic tails of the cosine rolloff impulse response, figure (5.29). These damped sinusoids cause instantaneous frequency fluctuations at the rate of $1/f_c$ per second, which is approximately equivalent to the introduction of new LHP analytic signal zeros at this rate. The resulting reconstruction spectrum distortion may be explained in terms of rate of LHP zero occurrence or by narrow band frequency modulation by a quasi-sinusoid.

The effect of lowpass filtering $\omega_i(t)$ is therefore twofold. Information is lost through the removal and attenuation of fluctuations corresponding to the zeros of $\Psi(z)$ and distortion is introduced by alteration of the "shape" of fluctuations which results in the introduction of new analytic signal zeros.

The observed intelligibilities of constant amplitude, lowpass instantaneous frequency reconstructions imply that considerable speech information is maintained by the instantaneous frequency waveform lowpass filtered to 1000Hz. The preserved information is now investigated in terms of the above discussion.

Vowel

The "basic" shape of the instantaneous frequency waveform of a vowel has been shown to be that generated by the major spectral components of the first formant (Section (4.3.1.3.)). This waveform does not generally exhibit wild fluctuations and its shape is well maintained by lowpass filtering to 1000Hz. The first formant of a vowel is therefore reasonably well reproduced by constant amplitude, lowpass instantaneous frequency reconstruction.

Second and third formant information corresponds to instantaneous frequency fluctuations around the "basic" waveforms at rates usually exceeding 1000Hz. Lowpass filtering removes or reduces the magnitudes of these fluctuations resulting in a loss of analytic signal zero information corresponding to the higher formants.

Some high frequency information may be preserved, however, if any of the superimposed fluctuations were caused by UHP analytic signal zeros. If the resulting lowpass instantaneous frequency rise retains sufficient magnitude, then the positions of the corresponding pair of zero crossings of the speech waveform are not greatly distorted in the reconstruction. In this way, higher formant information may be preserved by a constant amplitude, lowpass instantaneous frequency reconstruction.

If it is assumed that the process of differentiation does not significantly alter the first formant, but amplifies the second and third formants of a vowel, then the instantaneous frequency waveform of a differentiated vowel exhibits a higher proportion of UHP zero "rise" fluctuations about its "basic" waveform. Given sufficient lowpass instantaneous frequency bandwidth (usually $\geq$ 1000Hz) a constant amplitude, lowpass instantaneous frequency reconstruction of a differentiated vowel should therefore exhibit more high frequency information bearing zero crossings. The zero crossings of a real waveform are reproduced with less distortion than its complex zeros in a constant amplitude, lowpass instantaneous frequency reconstruction.

If the amplitudes of higher formants vary during the utterance of a vowel, the number of zero crossings per cycle reproduced by a lowpass instantaneous frequency reconstruction will also vary, resulting in "burbling".

Unvoiced Fricative

Figure (5.24) illustrates the narrowing of the pdf of the instantaneous frequency of an unvoiced fricative under lowpass filtering. Such a narrowing about $\overline{\omega_i(t)}$ is the

direct result of reduction of the amplitudes of instantaneous frequency fluctuations.

Lowpass instantaneous frequency reconstructions of unvoiced fricatives exhibit corresponding bandwidth reduction about $\overline{\omega_i(t)}$.

Clipping

The similarity between constant amplitude, lowpass instantaneous frequency reconstruction of differentiated speech and clipped differentiated speech is further illustrated by the instantaneous parameters of the clipped waveform.

The process of infinite peak clipping is often assumed to be one which strips a waveform of its amplitude function, but maintains phase data. Instantaneous parameter analysis of clipped baseband signals, however, reveals that it is generally UHP analytic signal zero positions which are preserved by clipping, and any LHP analytic signal zeros are either lost or converted to UHP (if an associated vector inner loop intersects the line $s(t)=0$). Lowpass filtering of the clipped waveform to its original bandwidth restores some amplitude information (dips) which corresponds to rises of $\omega_i(t)$ whose dynamic range exceed the lowpass cutoff frequency, f_c. The important instantaneous amplitude structure corresponding to UHP zeros whose associated rises of $\omega_i(t)$ do not exceed f_c (such as those which define the "basic" instantaneous waveforms of many vowels), is not reproduced by simple re-filtering of the clipped waveform.

CHAPTER 6

(6.1) AVENUES FOR FUTURE RESEARCH

The purpose of this section is to indicate two areas of current research which have not been fully explored and to suggest a possible direction for further work.

It was demonstrated in Chapter 5 that the "shape" of the impulse response used in FIR lowpass filtering of instantaneous frequency has a bearing on distortions exhibited by lowpass instantaneous parameter reconstructions. Such distortions might be reduced through the use of a more "suitable" impulse response and it is suggested that attempts be made to filter with impulse responses which do not possess the "periodic tails" exhibited by the cosine rolloff version.

Results from the previous chapter also suggest that the quality of lowpass filtered instantaneous parameter reconstructed speech may be improved by speech pre-processing. A major quality improvement would be afforded by the elimination of "burbling" distortion, which has been shown to be caused by LHP $\leftrightarrow$ UHP analytic signal zero conversions during the utterance of a voiced phoneme.

Analysis of particular phonemes has shown that these zero conversions usually occur during periods of low instantaneous amplitude (as dictated by the shape of the "basic" instantaneous waveforms of a vowel). It is possible that the rate of zero conversion may be lowered by reducing the severity of the periodic amplitude null. This is possible by selectively altering the phase of major spectral components of the vowels first formant, thereby destroying their almost in-phase relationship.

If the first formant of a hypothetical vowel is approximated by three harmonically related sinusoids, then the analytic signal which defines the "basic" instantaneous waveforms can be written

$$\Psi(z) = \sum_{n=3}^{5} A_n e^{j\phi_n} . e^{jn\omega z} \qquad \ldots (6.1)$$

where n is the harmonic number, A_n and ϕ_n the magnitude and phase of the n^{th} harmonic and ω the spectral line spacing. The zeros of the analytic function are defined by the polynomial

$$A_5.e^{j\phi 5}.e^{j5\omega z}+A_4.e^{j\phi 4}.e^{j4\omega z}+A_3.e^{j\phi 3}.e^{j3\omega z}=0 \quad \ldots (6.2)$$

Altering the phase of each sinusoid of the first formant model changes the complex coefficients of the polynomial, equation (6.2), and in doing so shifts the locations of the two non-removed zeros of $\Psi(z)$. Such a shift changes the shape of the vowels "basic" instantaneous waveforms and must therefore change the shape of the vowels instantaneous amplitude waveform. Processing based on this analysis may reduce "peakiness" of the instantaneous amplitude of a vowel, but would not be expected to affect intelligibility as the phonemes amplitude spectrum is unchanged.

The analytic signal model of the previous example shows how the first formant of a vowel may be defined by 2 or 3 non-removed complex zero positions, plus the rate of occurrence of removed zeros. As a vowel is almost stationary, these non-removed zeros would not be expected to shift significantly during its utterance and their positions are therefore parameters which could be suitable for coding in a low bit rate speech transmission scheme.

The analytic signal, $\Psi(z)$, which corresponds to the first formant or a sub-band of speech may be generated from its real and imaginary projections, $s(t)$ and $\hat{s}(t)$. Unfortunately, once obtained, factorisation of $\Psi(z)$ for its zero positions is not trivial.

Work with the instantaneous parameters of speech sub-bands has suggested one possible method for estimating non-removed analytic signal zero locations. If the sub-band

is of sufficiently narrow bandwidth that $\Psi(z)$ exhibits few non-removed zeros per cycle and if these zero locations are well separated in real time, the waveforms $a(t)$ and $\omega_i(t)$ may be used to estimate σ_n, the imaginary component of a particular non-removed zero.

It is known that the zeros of $a^2(t)$ are those of $\Psi(t)$ and $\Psi^*(t)$ and that they occur in complex conjugate pairs. If a complex conjugate pair of zeros of $a^2(t)$ are well separated in real time from the preceding and following pair, the resulting dip of $a^2(t)$ may be approximated by a quadratic curve. Given a complex conjugate pair of zeros at real time $\tau_n (z_n=\tau_n \pm j\sigma_n)$ which meet the above condition, then $a^2(t)$ in the proximity of τ_n is approximated by

$$c(t) = t^2-2t\tau_n+\tau_n{}^2+\sigma_n{}^2 \qquad \ldots (6.3)$$

At its minimum, $c(t)=\sigma_n{}^2$. The minimum of the corresponding dip of $a(t)$ is therefore $|\sigma_n|$ and $\mathrm{sgn}(\sigma_n)$ may be determined by reference to the polarity of the fluctuation of $\omega_i(t)$ at time τ_n.

A successful use of this technique may be illustrated using the curves of $a(t)$ in figures 2.15, 2.16 and 2.17. The value of the imaginary component of the non-removed analytic signal zeros in each case can be estimated from

$$\omega.|\sigma| = \min(a(t))/\max(a(t)) \qquad \ldots (6.4)$$

which obtains $\omega.|\sigma| = 0.61$, 0.33 and 0.05 respectively for figures 2.15.4, 2.16.4 and 2.17.4. These compare favourably with the actual values.

Reliable estimation of σ can only be expected if analytic signal zeros are well separated in real time (i.e. dips of $a(t)$ do not superimpose). It may be possible to ensure this for a sub-band of speech by pre-processing of signal phase as suggested previously for reducing the peakiness of $a(t)$.

(6.2) CONCLUSION

The instantaneous parameters of analytic vector representations of speech and sub-bands of speech have been generated and investigated. The characteristics of these waveforms have been related to characteristics of the real signal time waveform, vector loci and the spectral and zero structure of the associated analytic signal.

As parameters of the speech signal, both a(t) and $\omega_i(t)$ carry pitch and "voiced/unvoiced decision" information and combine to define the speech amplitude and phase spectrum. Processed instantaneous parameters, such as those transmitted by frequency division vocoders, maintain pitch and voiced/unvoiced information, but the spectral envelope of synthesised speech may suffer distortion (possibly resulting in "burbling"). The fact that the instantaneous parameters of sub-bands of speech may exhibit unlimited bandwidth and dynamic range suggests that they are grossly distorted by some frequency division vocoder systems.

Real time generation of the instantaneous parameters of telephone bandwidth speech has been achieved, but shown to require sampling rates in excess of 10,000 per second and a 12 bit digital word length (to ensure accuracy in calculating $\omega_i(t)$). The associated real time vector loci exhibit reasonably repeatable "shapes", but emphasise slight aperiodicities of voiced phonemes. Chart recordings of the time averaged instantaneous frequency of voiced phonemes show it to be an estimator of first formant frequency and illustrate similarities between $\overline{\omega_i(t)}$ and zero crossing rate.

Notable features of the instantaneous parameters of vowels are the average instantaneous frequency, "basic" waveforms and superimposed fluctuations. Average instantaneous frequency over one cycle of a vowel does not necessarily coincide with the frequency of a major spectral

line of the first formant, but is always a multiple of the spectral line spacing frequency and is on or just above the first formant frequency.

The "basic" instantaneous parameter waveshapes are those generated by the major spectral components of the first formant. Superimposed fluctuations around these waveshapes are caused by the spectral components of upper formants and the rate of superimposed fluctuation occurrence is related to formant difference frequencies.

The pdfs of the instantaneous parameters of band-limited unvoiced phonemes are reasonably accurately approximated by those of bandpass filtered Gaussian noise. As is the case with narrow band noise, these phonemes are well defined by the statistics of the pdf of their instantaneous frequency. Important features of this pdf are the centre frequency ($\overline{\omega_i(t)}$) and probability $p\{\omega_i(t)=\overline{\omega_i(t)}\}$. From this, an equivalent rectangular noise bandwidth may be calculated. The instantaneous frequency curve of the unvoiced plosive /t/ in figure 4.35 shows it to be a useful form of spectral analysis for such dynamic phonemes.

The instantaneous parameters of voiced fricatives may be described as hybrids, being related to those of vowels and unvoiced fricatives. The shapes of pdfs of the instantaneous amplitude and frequency of voiced fricatives resemble those of noise or a sinusoid plus noise at different times during one cycle of the phoneme. The amplitude spectra of unvoiced fricatives show them to consist of low frequency voiced energy plus high frequency noise.

The pdf of the instantaneous frequency of a voiced phoneme does not reflect the phonemes spectral envelope. The peak of the pdf, however, is usually at a much lower frequency than that of an unvoiced phoneme, and this is clearly illustrated by instantaneous parameter analyses of words. Onset of unvoiced phonemes and transitions between

voiced and unvoiced phonemes may be accurately located by tracking the peak of a dynamic pdf of instantaneous frequency ($n \simeq 100$ for $\omega_i(t)$ sampled at 10,000 samples per second).

Assuming that original instantaneous parameters are recorded with sufficient bandwidth and dynamic range, distortion resulting from modified instantaneous parameter reconstruction can be described in terms of amplitude spectrum distortion. This often involves the introduction of "image" spectral components whose locations may be predicted from knowledge of the average instantaneous frequency value.

Distortions inherent in constant amplitude and clipped versions of speech can be explained in terms of "image" components. Instantaneous parameter analysis of clipped waveforms, however, provides a method of determining the information (in terms of zeros of the analytic signal) retained after clipping.

Lowpass instantaneous parameter reconstruction of speech has revealed that instantaneous frequency is more important to reconstruction intelligibility than instantaneous amplitude. Although the use of a wideband version of $a(t)$ improves reconstruction quality, telephone bandwidth speech maintains its intelligibility if it is reconstructed using a version of $\omega_i(t)$ not lowpass filtered to below $f_c \simeq 1000$Hz. This cutoff frequency may, however, be slightly dependent on the "shape" of the lowpass filter impulse response, $h(t)$, and the above value of f_c applies particularly to the filter described in figure 5.29.

In general, lowpass instantaneous parameter reconstruction of voiced phonemes results in a loss of high frequency information and the introduction of "burbling". Lowpass instantaneous parameter reconstruction of unvoiced phonemes results in narrowing of the bandwidth around their centre frequency ($\overline{\omega_i(t)}$).

The way in which speech information is coded by its instantaneous parameters clarifies the limitations of a transmission system based on lowpass instantaneous parameter reconstruction. Communications quality speech has been obtained in reconstruction using instantaneous parameters passed through a total equivalent analogue bandwidth of approximately 1000Hz. Assuming that the lowpass instantaneous parameters require quantization to 8 bits accuracy, this corresponds to speech transmission at 16 kbits per second. Further bit rate reduction may be possible, however, by intelligently quantizing each parameter according to its particular characteristics. Phase vocoders claim communications quality output at approximately 10 kbits per second.

The paper currently in preparation describes a method for the reduction of the peak/rms power ratio of speech, based on "phase scrambling" of a vowel's first formant to alter the "basic" instantaneous amplitude waveform. For the case of a three sinusoid first formant model, the least peaky instantaneous amplitude waveform is obtained when the two lower frequency sinusoids are in phase and the high frequency component is 180° out of phase. This condition also corresponds to the two non-removed analytic signal zeros being separated by π/ω seconds, where ω is the spectral line spacing.

The conclusion drawn from the paper is that the least peaky instantaneous amplitude waveforms correspond to analytic signals whose non-removed zeros are distributed evenly in real time. Although blind "phase scrambling" of spectral components does not guarantee a perfectly even distribution of non-removed zeros, it can be expected to spread the characteristically bunched zeros corresponding to the first formant of a vowel thus reducing the overall peak/rms power ratio.

APPENDIX A

DERIVATION OF FORMULAE FOR THE INSTANTANEOUS PARAMETERS OF A TWO SINUSOID SIGNAL

Equation (2.23) for the real signal is

$$s(t)=A\cos\omega_1 t+B\cos(\omega_2 t+\theta).$$

The Hilbert Transform is therefore

$$\hat{s}(t)=A\sin\omega_1 t+B\sin(\omega_2 t+\theta).$$

INSTANTANEOUS AMPLITUDE

The instantaneous amplitude signal is defined

$$a(t)=\{s^2(t)+\hat{s}^2(t)\}^{\frac{1}{2}}.$$

Expanding this in terms of s(t) and $\hat{s}(t)$

$$\begin{aligned}a(t)&=\{(A\cos\omega_1 t+B\cos(\omega_2 t+\theta))^2+(A\sin\omega_1 t+B\sin(\omega_2 t+\theta))^2\}^{\frac{1}{2}}\\&=\{A^2\cos^2\omega_1 t+2AB\cos\omega_1 t.\cos(\omega_2 t+\theta)+B^2\cos^2(\omega_2 t+\theta)\\&\quad+A^2\sin^2\omega_1 t+2AB\sin\omega_1 t.\sin(\omega_2 t+\theta)+B^2\sin^2(\omega_2 t+\theta)\}^{\frac{1}{2}}\\&=\{A^2+B^2+2AB\cos((\omega_2 t+\theta)-\omega_1 t)\}^{\frac{1}{2}}\\&=\{A^2+B^2+2AB\cos(\omega_d t+\theta)\}^{\frac{1}{2}}\end{aligned}$$

ω_d is the difference frequency

$$\omega d=\omega_2-\omega_1$$

and the above result corresponds to equation (2.24).

INSTANTANEOUS PHASE

The instantaneous phase is defined

$$\phi(t)=\tan^{-1}\{\hat{s}(t)/s(t)\}$$

Expanding in terms of $s(t)$ and $\hat{s}(t)$

$$\phi(t)=\tan^{-1}\left[\frac{A\sin\omega_1 t+B\sin(\omega_2 t+\theta)}{A\cos\omega_1 t+B\cos(\omega_2 t+\theta)}\right]$$

Using the definition of ω_d

$$\omega_2 t+\theta=\omega_1 t+\omega_d t+\theta$$

The equation for $\phi(t)$ can now be rewritten

$$\phi(t)=\tan^{-1}\left[\frac{A\sin\omega_1 t+B\sin\omega_1 t\cos(\omega_d t+\theta)+B\cos\omega_1 t\sin(\omega_d t+\theta)}{A\cos\omega_1 t+B\cos\omega_1 t\cos(\omega_d t+\theta)-B\sin\omega_1 t\sin(\omega_d t+\theta)}\right]$$

$$=\tan^{-1}\left[\frac{A\sin\omega_1 t\{1+{}^{B}/_{A}\cos(\omega_d t+\theta)\}+B\cos\omega_1 t.\sin(\omega_d t+\theta)}{A\cos\omega_1 t\{1+{}^{B}/_{A}\sin(\omega_d t+\theta)\}-B\sin\omega_1 t.\sin(\omega_d t+\theta)}\right]$$

$$=\tan^{-1}\left[\frac{\tan\omega_1 t+{}^{B}/_{A}\sin(\omega_d t+\theta)/(1+{}^{B}/_{A}\cos(\omega_d t+\theta))}{1-({}^{B}/_{A}\sin(\omega_d t+\theta)/(1+{}^{B}/_{A}\cos(\omega_d t+\theta)))\tan\omega_1 t}\right]$$

$$=\tan^{-1}\left[\tan\left[\omega_1 t+\tan^{-1}\left[\frac{{}^{B}/_{A}\sin(\omega_d t+\theta)}{1+{}^{B}/_{A}\cos(\omega_d t+\theta)}\right]\right]\right]$$

$$=\omega_1 t+\tan^{-1}\left[\frac{{}^{B}/_{A}\sin(\omega_d t+\theta)}{1+{}^{B}/_{A}\cos(\omega_d t+\theta)}\right]$$

This result corresponds to equation (2.25).

INSTANTANEOUS FREQUENCY

Equation (2.26) for instantaneous frequency is obtained by differentiating equation (2.25) with respect to time

$$\omega_i(t) = {}^{d}/_{dt}\left[\omega_1 t + \tan^{-1}\left[\frac{{}^{B}/_{A}\sin(\omega_d t+\theta)}{1+{}^{B}/_{A}\cos(\omega_d t+\theta)}\right]\right]$$

$$= \omega_1 + {}^{d}/_{dt}\left[\tan^{-1}\left[\frac{{}^{B}/_{A}\sin(\omega_d t+\theta)}{1+{}^{B}/_{A}\cos(\omega_d t+\theta)}\right]\right]$$

It is known that

$${}^{d}/_{dt}\tan^{-1}(\mu) = (1/(1+\mu^2)).{}^{d\mu}/_{dt}.$$

Making the substitutions

$$\mu = \frac{{}^{B}/_{A}\sin(\omega_d t+\theta)}{1+{}^{B}/_{A}\cos(\omega_d t+\theta)}$$

and $${}^{d\mu}/_{dt} = \frac{\omega_d({}^{B^2}/_{A^2}) + \omega_d({}^{B}/_{A})\cos(\omega_d t+\theta)}{\{1+({}^{B}/_{A})\cos(\omega_d t+\theta)\}^2}$$

allows us to write

$$\omega_i(t) = \omega_1 + \frac{\omega_d({}^{B^2}/_{A^2}) + \omega_d({}^{B}/_{A})\cos(\omega_d t+\theta)}{1+({}^{B^2}/_{A^2}) + 2({}^{B}/_{A})\cos(\omega_d t+\theta)}$$

This is equation (2.26) for instantaneous frequency.

APPENDIX B

FOURIER ANALYSIS OF THE INSTANTANEOUS PARAMETERS OF A TWO SINUSOID SIGNAL

INSTANTANEOUS AMPLITUDE

Rewriting equation (2.24) for instantaneous amplitude in terms of ρ and ξ yields

$$a(t) = A(1+\rho^2+2\rho\cos(\xi))^{\frac{1}{2}}$$

By means of a binomial expansion, the factor $(1+\rho^2+2\rho\cos(\xi))^{\frac{1}{2}}$ can be rewritten

$$a_0+a_1\cos(\xi)+a_2\cos(2\xi)+\ldots$$

where the first two coefficients are

$$a_0=1+\rho^2/4+\rho^4/64+\rho^6/256+\ldots$$

$$a_1=\rho(1-\rho^2/8-\rho^6/64-5\rho^6/1024-\ldots)$$

The general expression for a_n is

$$a_n=2(-1)^n\left[\frac{1(3)\ldots(2n-1)}{n!}\right]\frac{\rho^n}{2^n}\left[\frac{-1}{2n-1}+\frac{1}{1(n+1)}\cdot\frac{\rho^2}{2!}\right.$$

$$\left.+\sum_{k=2}^{\infty}\frac{1(3)\ldots(2k-3)}{\cdot\ k!2^{2k}}\cdot\frac{(2n+1)(2n+3)\ldots(2n+2k-3)}{(n+1)(n+2)\ldots(n+k)}\rho^{2k}\right]$$

(Ref. 74)

INSTANTANEOUS FREQUENCY

The variable term of equation (2.26) for instantaneous frequency can be written

$$v=\omega_d\left[\frac{\rho^2+\rho\cos(\xi)}{1+\rho^2+2\rho\cos(\xi)}\right]$$

When ρ is restricted to the range $-1\leq\rho\leq1$, v is proportional to the Fourier series

$$v \propto -\omega_d \sum_{n=1}^{\infty} (-\rho)^n\cos(n\xi)$$

(Ref. 74)

APPENDIX C

CALCULATION OF ZERO POSITIONS FOR EXAMPLE (2.5.1)

By equation (2.67), the non-removed analytic signal zeros corresponding to the real signal of example (2.5.1) are defined by

$$z=-j\ln\left|1/\sqrt{B}\right|/\omega \text{ or } z=\pi/\omega-j\ln\left|1/\sqrt{B}\right|/\omega$$

In both cases the imaginary component is $\sigma_c=\ln\left|1/\sqrt{B}\right|/\omega$.

The real signal itself always exhibits a pair of zero crossings at $\tau=\pi/2\omega$ and $\tau=3\pi/2\omega$. The remaining four zeros per cycle may be conjugate pairs at $\tau=0$ and $\tau=\pi/\omega$ or distinct zero crossings according to the solution of

$$e^{j\omega z}=\pm\left[\frac{(1+B)\pm\sqrt{(-3B-1)(B-1)}}{2B}\right]^{\frac{1}{2}} \qquad \ldots \text{ (C.1)}$$

B=0.25

$$\sigma_c=-\ln\left|1/\sqrt{.25}\right|/\omega=-0.69/\omega$$

The solutions of equation (C.1) for B=0.25 are

$$e^{j\omega z}=\pm 2.18 \text{ and } \pm 0.46$$

which maps to the complex time plane as

$$z=\pm j0.78/\omega \text{ and } z=\pi/\omega \; j0.78/\omega$$

B=0.5

$$\sigma_c=-\ln\left|1/\sqrt{0.5}\right|/\omega=-0.35/\omega$$

The solutions of equation (C.1) for B=0.5 are

$$e^{j\omega z} = \pm 1.62 \text{ and } \pm 0.62$$

so $z = \pm j0.48/\omega$ and $z = \pi/\omega \pm j0.48/\omega$

B=0.9

$$\sigma_c = -\ln|1/\sqrt{0.9}|/\omega = -0.05/\omega$$

The solutions of equation (C.1) for B=0.9 are

$$e^{j\omega z} = \pm 1.19 \text{ and } \pm 0.84$$

so $z = \pm j0.17/\omega$ and $z = \pi/\omega \pm j0.17/\omega$

B=2

$$\sigma c = -\ln|1/\sqrt{2}|/\omega = +0.35/\omega$$

The solutions of equation (C.1) for B=2 are

$$e^{j\omega z} = \pm e^{j\pm 0.12\pi}$$

so $z = \pm 0.12\,\pi/\omega$ and $z = (\pi \pm 0.12\pi)/\omega$.
In this case, equation (C.1) locates 4 distinct zero crossings.

APPENDIX D

CIRCUITRY DETAILS OF HARDWARE ANALYTIC DECODER

D.1 Speech channel pre-filter.

D.2 TAD FIR Hilbert transformer.

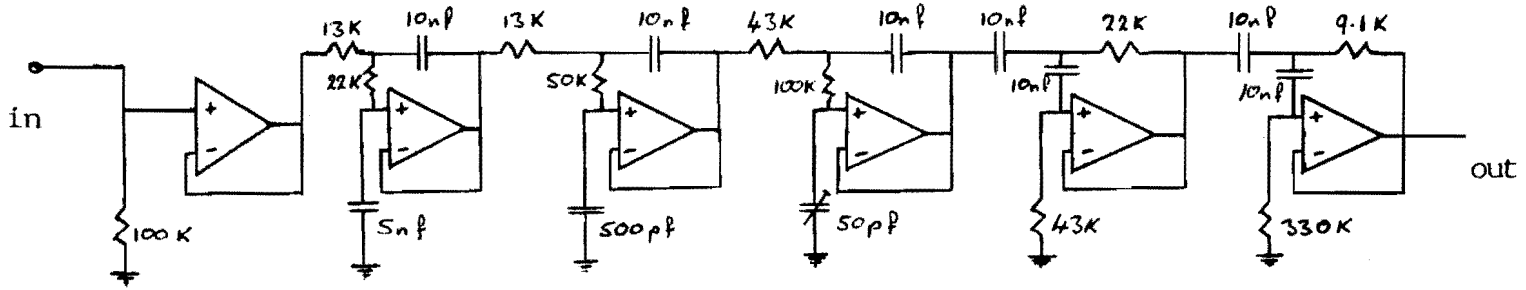
D.3 Matched smoothing filters.

D.4 Data acquisition.

D.5 Instantaneous amplitude.

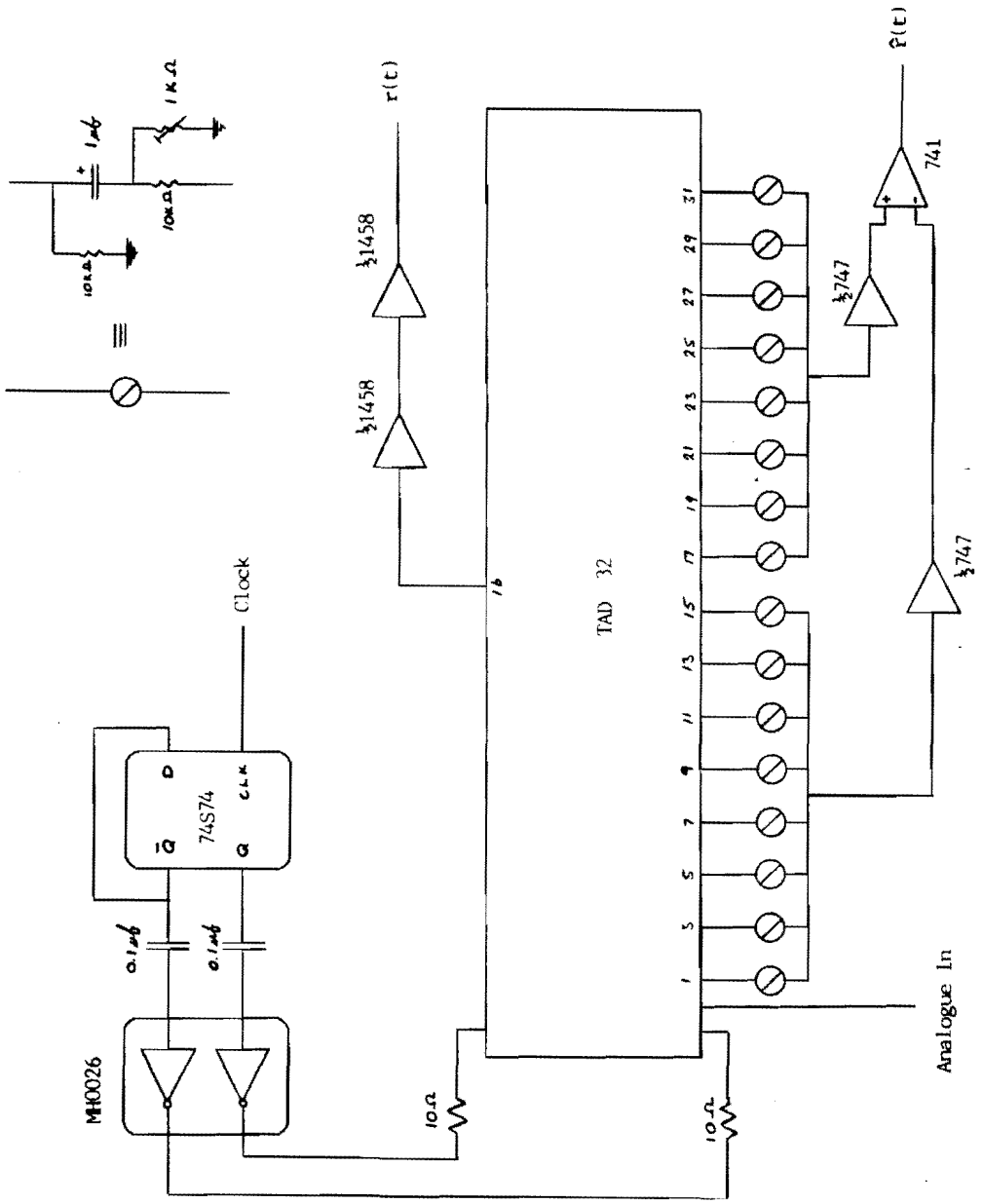
D.6 Instantaneous phase.

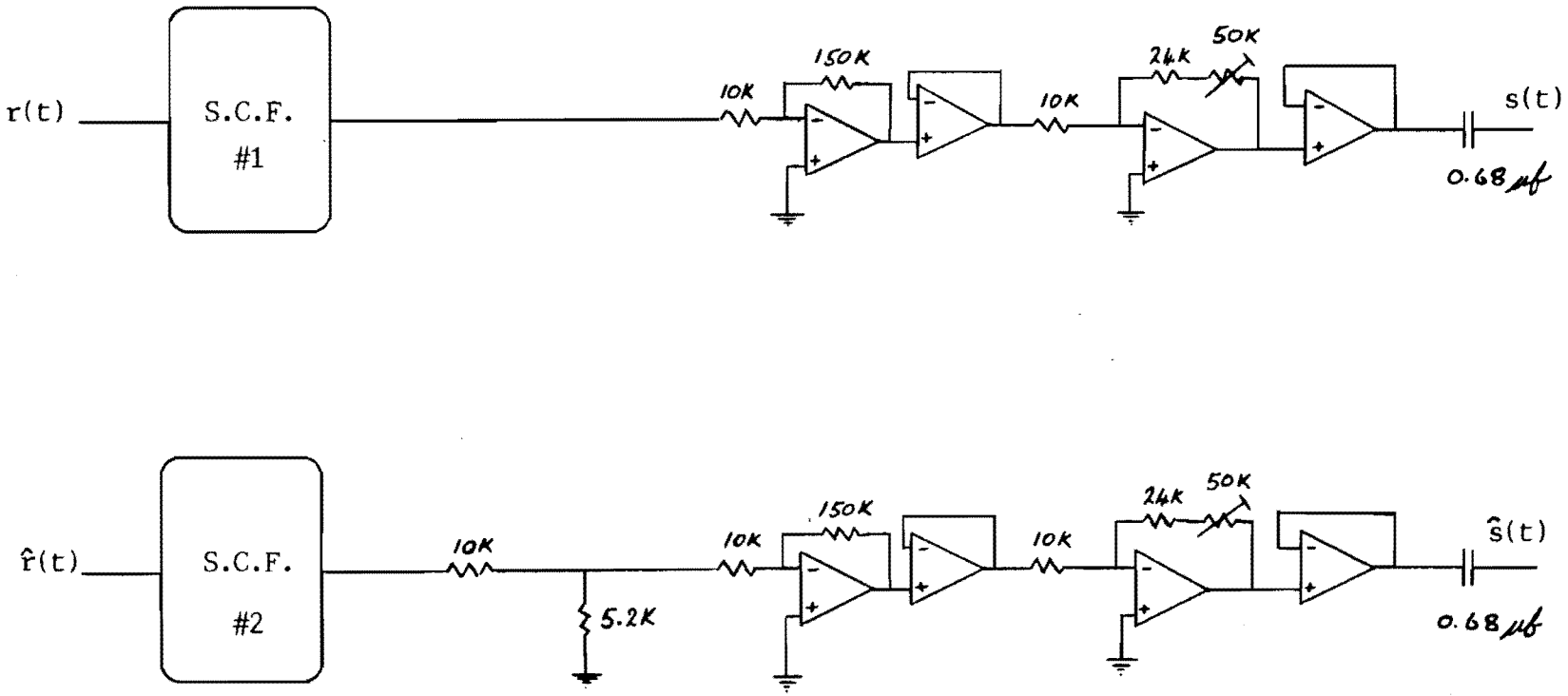
D.7 Instantaneous frequency.



All Operational Amplifiers Are LM741

D.2 TAD FIR HILBERT TRANSFORMER

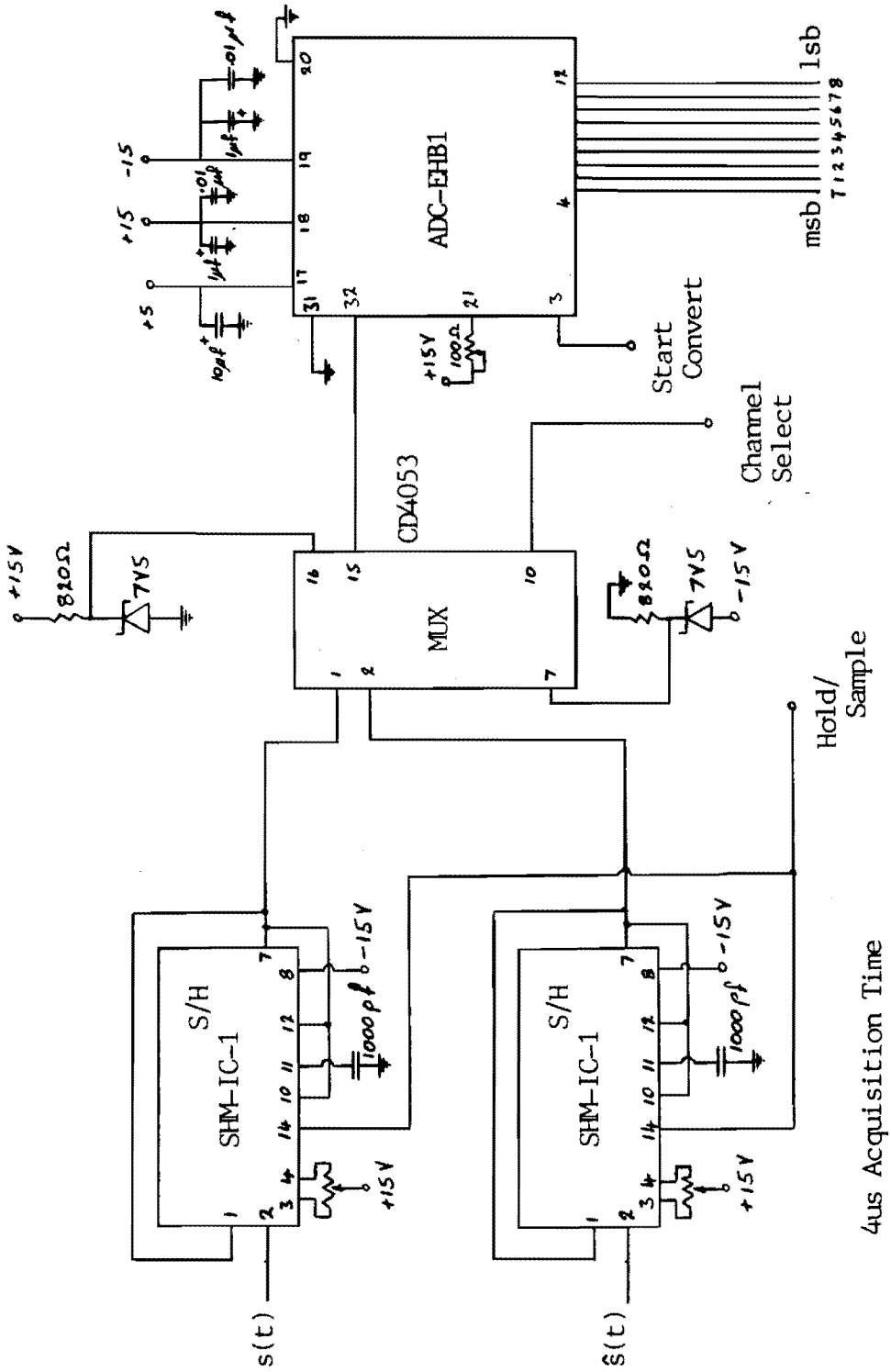




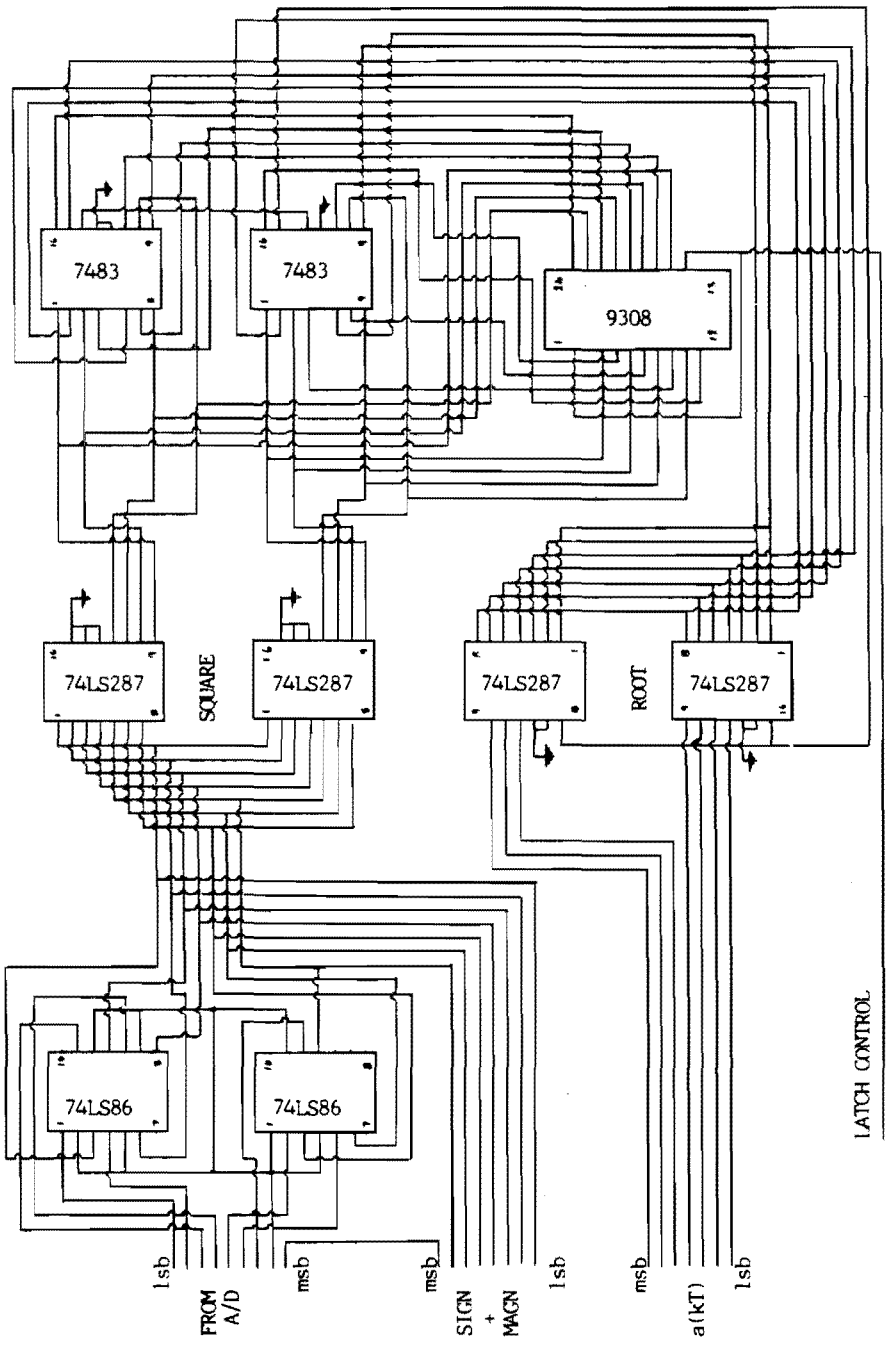
S.C.F. : Speech Channel Pre-filter

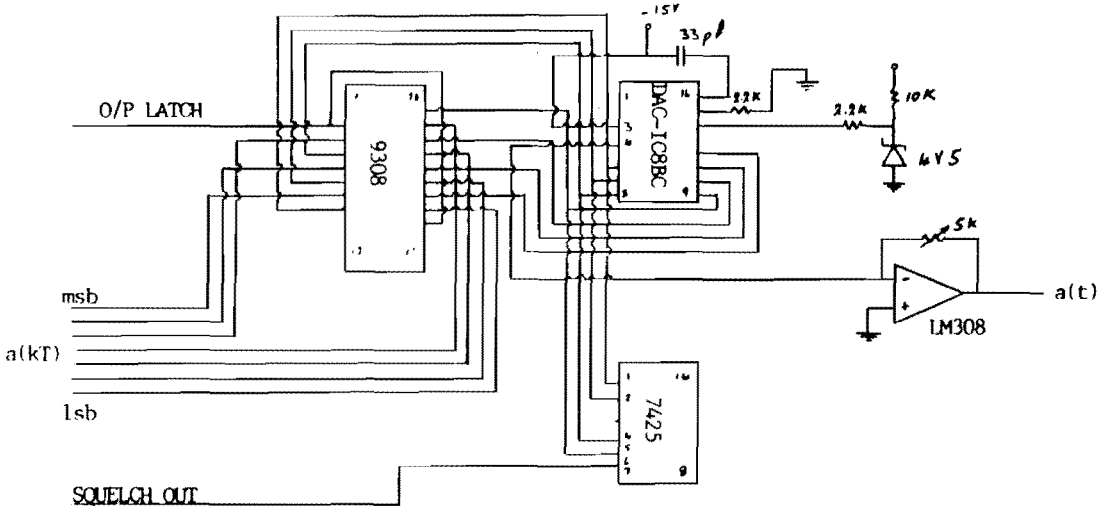
All Operational Amplifiers CA3140

D.4 DATA ACQUISITION

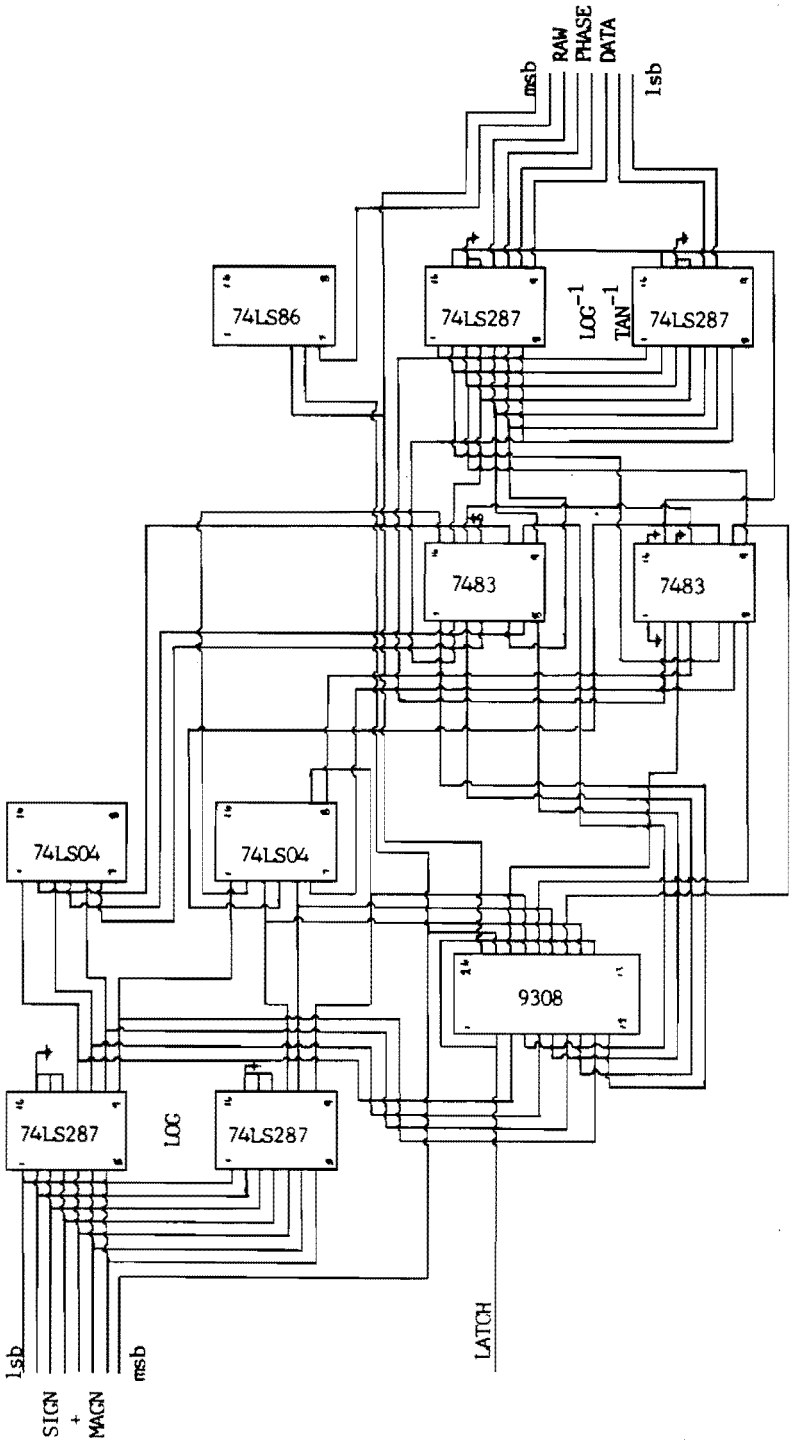


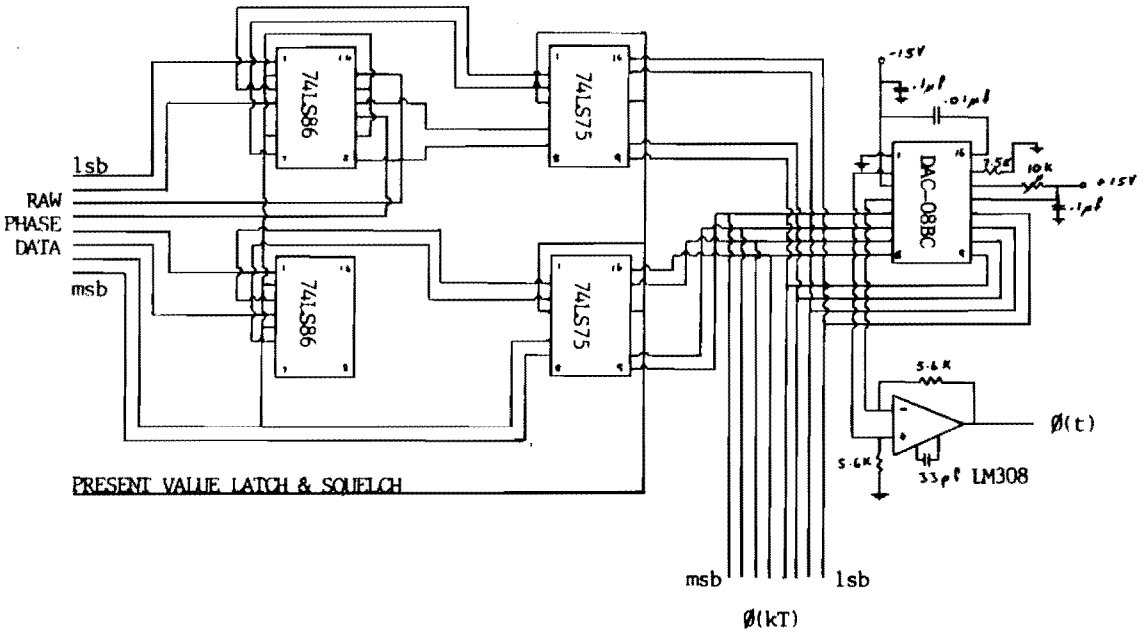
D.5 INSTANTANEOUS AMPLITUDE



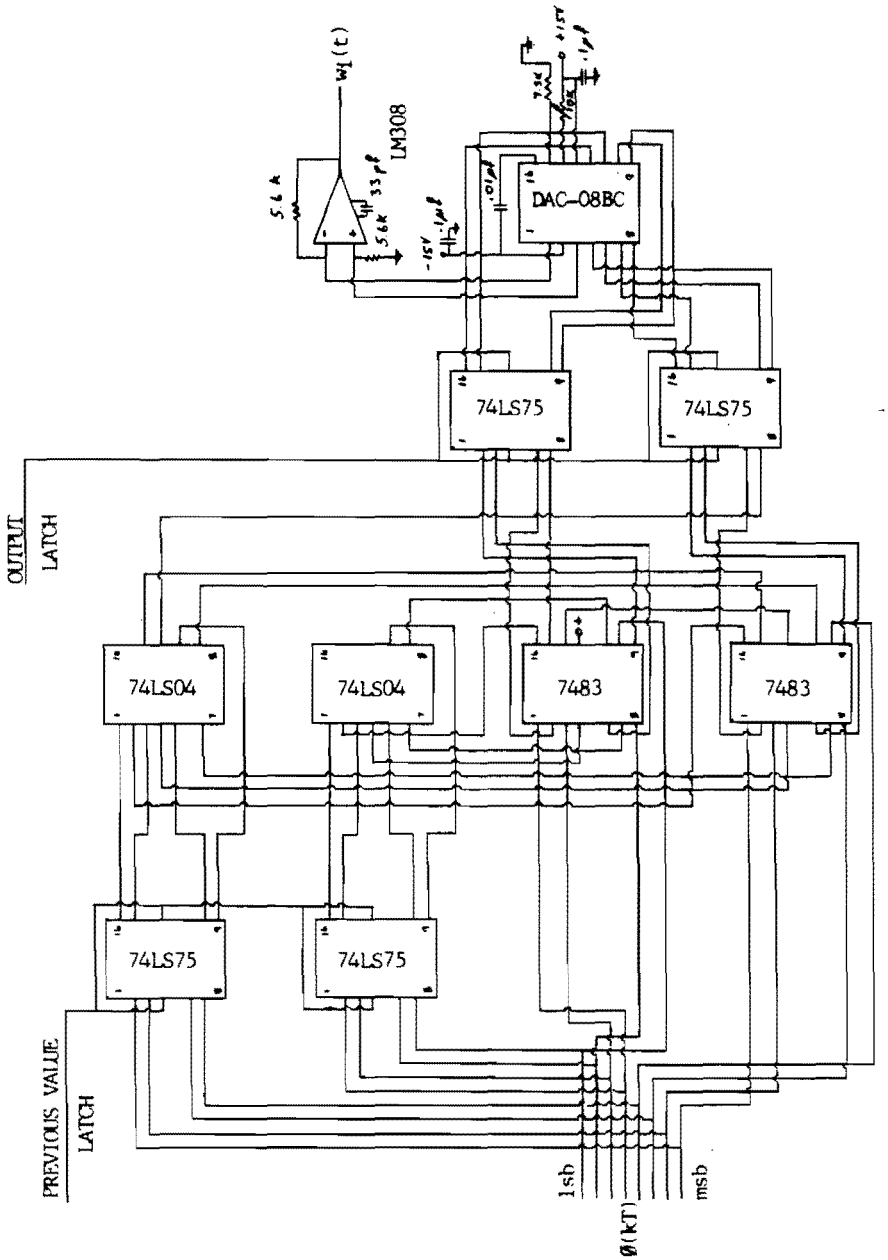

O/P LATCH
msb
a(kT)
lsb
SQUELCH OUT
9308
DAC-IC8BC
7425
LM308
a(t)
-15V
33pf
22K
2.2K
10K
4V5
5k

D.6 INSTANTANEOUS PHASE




lsb
RAW
PHASE
DATA
msb
74LS86
74LS86
74LS75
74LS75
DAC-08BC
-15V
+15V
10K
5.6K
5.6K
33pf
LM308
Ø(t)
PRESENT VALUE LATCH & SQUELCH
msb
lsb
Ø(kT)

D.7 INSTANTANEOUS FREQUENCY



REFERENCES

1. FLANAGAN, J.L. "Speech Analysis Synthesis and Perception" Springer-Verlag 1972.

2. LEVITT; PICKETT; HOUDE (Editors) "Sensory Aids for the Hearing Impaired" I.E.E.E. Press 1980.

3. FANT, G. "The Acoustics of Speech" Proceedings 3rd International Congress on Acoustics pp 188-201 1959.

4. MILLER, R.L. "Nature of the Vocal Cord Wave' J.A.S.A. Vol - 31 pp 667-677 June 1959.

5. FLANAGAN, J.L. "Voices of Men and Machines" J.A.S.A. Vol - 51 pp 1375-1387 March 1972.

6. RABINER, L.A.; SCHAFER, R.W. "Digital Processing of Speech Signals" Prentice-Hall 1978.

7. RICHARDS, D.L. "Telecommunications by Speech" Butterworth and Co. London 1973.

8. NGUYEN, D.T.; GUERIN, B. "Effects of Nasal Coupling on the Vowels" Paper presented at 99th meeting of the Acoustical Society of America, Atlanta 1980.

9. FUJIMURA, O. "Analysis of Nasal Consonants" J.A.S.A. Vol - 34 pp 1865-1875 December 1962.

10. HUGHES, G.W.; HALLE, M "Spectral Properties of Fricative Consonants" J.A.S.A. Vol - 28, No.2 pp 303-310 March 1956.

11. HEINZ, J.M.; STEVENS, K.N. "On the Properties of Voiceless Fricative Consonants" J.A.S.A. Vol - 33 pp 589-596 May 1961.

12. HUGHES, G.W.; RADLEY, J.P.A. "Acoustic Properties of Stop Consonants" J.A.S.A. Vol - 29, No.1 January 1957.

13. SCHROEDER, M.R. "Models of Hearing" Proceedings I.E.E.E. Vol - 36, No.9 September 1975.

14. SHAW, E.A.G. "The External Ear" contained in "Handbook of Sensory Physiology" Vol - V/1, Auditory System Edited by KEIDEL, W.D. and NEFF, W.D. Springer-Verlag 1974.

15. ELREDGE, E.H. "Inner Ear Cochlear Mechanics and Cochlear Potentials" contained in "Handbook of Sensory Physiology" Vol - V/1, Auditory System Edited by KEIDEL, W.D. and NEFF, W.D. Springer-Verlag 1974.

16. VON BEKESY, G. "Experiments in Hearing" McGraw-Hill 1960.

17. DAVID, E.E.; MILLER, J.E.; MATHEWS, M.V. "Monaural Phase Effects in Speech Perception" Proceedings 3rd International Conference on Acoustics pp 227-229 September 1959.

18. PLOMP, R.; STEENEKEN, H.J.M. "Effects of Phase on the Timbre of Complex Tones" J.A.S.A. Vol - 46, No.2 pp 409-421 1969.

19. LADEFOGED, P.; BROADBENT, D.E. "Information Conveyed by Vowels" J.A.S.A. Vol - 29, No.1 pp 98-104 January 1957.

20. SHANNON, C.E.; WEAVER, W. "The Mathematical Theory of Communication" Urbana : University of Illinois 1949.

21. DUDLEY, H. "The Vocoder" Bell Lab Record Vol - 18 pp 122-126 1939.

22. GOLD, B.; RADER, C.M. "The Channel Vocoder" I.E.E.E. Trans on Audio and Acoustics Vol AU - 15, No.4 pp 148-161 December 1967.

23. SCHROEDER, M.R. "Vocoders" Proceedings I.E.E.E. Vol - 54 pp720-734 May 1966.

24. PETERSON, E.; COOPER, F.S. "Peakpicker : A Bandwidth Compression Device" J.A.S.A. Vol - 29 p777 (A) June 1957.

25. GOLD, B. "Techniques for Speech Bandwidth Compression Using Combinations of Channel Vocoders and Formant Vocoders" J.A.S.A. Vol - 38 pp2-10 1965.

26. FLANAGAN, J.; HOUSE, A.S. "Development and Testing of a Formant - Coding Speech Compression System" J.A.S.A. Vol - 28, No.6 pp1099-1106 November 1956.

27. DUDLEY, H. "Phonetic Pattern Recognition Vocoder for Narrow-Band Speech Transmission" J.A.S.A. Vol - 30, No.8 pp733-739 August 1958.

28. GRASSOT, F.; VOILLAUME, J.C. "Low Bit Rate Speech Transmission" Electrical Communication Vol - 55, No.4 pp316-322 1980.

29. MAITRA, S.; DAVIS, C.R. "A Speech Digitizer at 2400 Bits Per Second" I.E.E.E. Trans on Acoustics, Speech and Signal Processing Vol ASSP-27, No.6 pp729-733 December 1979.

30. MORGAN, D.R.; CRAIG, S.E. "Real Time Adaptive Linear Prediction Using the Least Mean Square Gradient Algorithm" I.E.E.E. Trans on Acoustics, Speech and Signal Processing Vol ASSP-24, No.6 pp494-507 December 1976.

31. MAKHOUL, J. "Linear Prediction : A Tutorial Review" Proceedings I.E.E.E. Vol - 63 pp561-580 April 1975.

32. ATAL, B.S.; HANAUER, S.L. "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave" J.A.S.A. Vol - 50 pp637-655 August 1971.

33. ATAL, B.S. "Automatic Speaker Recognition Based On Pitch Contours" J.A.S.A. Vol - 52 pp1687-1697 1973.

34. ROSENBERG, A.E.; SAMBUR, M.R. "New Techniques for Automatic Speaker Verification" I.E.E.E. Trans on Acoustics, Speech and Signal Processing Vol ASSP-23 pp169-176 April 1975.

35. GOLD, B.; RABINER, L.R. "Parallel Processing Techniques for Estimating Pitch Periods of Speech In the Time Domain" J.A.S.A. Vol - 46 pp442-448 August 1969.

36. MILLER, N.J. "Pitch Perception by Data Reduction" I.E.E.E. Trans on Acoustics, Speech and Signal Processing Vol ASSP-23 pp72-79 February 1975.

37. TUCKER, W.H.; BATES, R.H.T. "A Pitch Estimation Algorithm for Speech and Music" I.E.E.E. Trans on Acoustics, Speech and Signal Processing Vol - ASSP-26, No.6 December 1978.

38. RABINER, L.R. "On the Use of Autocorrelation Analysis for Pitch Detection" I.E.E.E. Trans on Acoustics Speech and Signal Processing Vol ASSP-25, No.1 February 1977.

39. SCHROEDER, M.R. "Parameter Estimation in Speech : A Lesson in Unorthodoxy" Proceedings I.E.E.E. Vol - 58, No.5 pp707-712 May 1976.

40. NOLL, A.M. "Cepstrum Pitch Determination" J.A.S.A. Vol - 41 pp293-309 February 1967.

41. RABINER, L.R.; CHENG, M.J. "A Comparitive Study of Several Pitch Detection Algorithms" I.E.E.E. Trans on Acoustics, Speech and Signal Processing Vol ASSP-24 pp399-417 October 1976.

42. ATAL, B.S.; RABINER, L.R. "A Pattern Recognition Approach to Voiced - Unvoiced - Silence Classification with Applications to Speech Recognition" I.E.E.E. Trans on Acoustics, Speech and Signal Processing Vol ASSP-24, No.3 pp201-212 June 1976.

43. SIEGEL, L.J.; BESSEY, A.C. "Voiced/Unvoiced/Mixed Classification of Speech" I.E.E.E. Trans on Acoustics, Speech and Signal Processing Vol ASSP-30 No.3 pp451-460 June 1982.

44. GOLD, B. "Experiments with Speech-like Phase in a Spectrally Flattened Pitch - Excited Channel Vocoder" J.A.S.A. V - 36, No.10 pp1892-1894 October 1964.

45. BELLAMY, J.C. "Digital Telephony" J. Wiley and Sons 1982.

46. OLIVER, B.M.; PIERCE, J.R.; SHANNON, C.E. "The Philosophy of P.C.M." Proceedings I.R.E. pp1324-1331 November 1948.

47. SCHINDLER, H.R. "Delta Modulation" I.E.E.E. Spectrum pp69-78 October 1970.

48. SCHILLING, D.L.; GARDONICK, J.; VANG, H.A. "Voice Encoding for the Space Shuttle Using Adaptive Delta Modulation" I.E.E.E. Trans on Communications Vol COM - 26, No.11 pp1652-1659 November 1978.

49. FLANAGAN, J.L.; SCHROEDER, M.R.; ATAL, B; CHROCHIERE, R.E.; JAYANT, N.S.; TRIBOLET, J.M. "Speech Coding" I.E.E.E. Trans on Communications Vol COM - 27, No.4 pp710-737 April 1979.

50. CHROCHIERE, R.E.; WEBBER, S.A.; FLANAGAN, J.L. "Digital Encoding of Speech in Sub-Bands" Proceedings of I.E.E.E. Conference on Acoustics, Speech and Signal Processing pp233-236 1976.

51. CHROCHIERE, R.E.; SAMBUR, M.R. "A Variable Band Coding Scheme for Speech Encoding at 4.8 kb/s" Proceedings I.E.E.E. Conference on Acoustics, Speech and Signal Processing pp444-447 1977.

52. BOGERT, B.P. "The Vobanc - A Two-to-One Speech Bandwidth Reduction System" J.A.S.A. Vol - 28, No.3 pp399-404 May 1956.

53. DAGEUT, J.L. "Speech Compression CODIMEX System" I.E.E.E. Trans on Audio Vol AU-11 pp63-70 March-April 1963.

54. SCHROEDER, M.R.; FLANAGAN, J.L.; LUNDRY, E.A. Bandwidth Compression of Speech by Analytic Signal Rooting" Proceedings I.E.E.E. Vol - 55, No.3 pp396-401 March 1967.

55. BOGNER, R.E. "Frequency Division in Speech Bandwidth Reduction" I.E.E.E. Trans on Communications Vol COM-13, No.4 pp438-451 December 1965.

56. GOLD, B.; RADER, C.M. "Systems for Compressing the Bandwidth of Speech" I.E.E.E. Trans on Audio and Electro Acoustics Vol AU-15, No.3 September 1967.

57. FLANAGAN, J.L.; GOLDEN, R.M. "Phase Vocoder" B.S.T.J. Vol - 45 pp1493-1509 November 1966.

58. SCHAFER, R.W.; RABINER, L.R. "Application of Digital Signal Processing to the Design of a Phase Vocoder Analyser" I.E.E.E. Conference on Speech Communication and Processing pp52-55 1972.

59. PORTNOFF, M.R. "Implementation of the Digital Phase Vocoder using the Fast Fourier Transform" I.E.E.E. Trans on Acoustics, Speech and Signal Processing Vol ASSP-24 pp243-248 June 1976.

60. CASTELLINO, P. "Speech Coding at Low Bit Rates : Quality Problems" Cselt Rapporti Tecnici Vol lX, No.5 Octobre 1981.

61. DECINA, M. "Managing I.S.D.N. Through International Standards Activities" I.E.E.E. Communications Magazine ppl9-25 September 1982.

62. Extract from the Report of the Working Party XVIII/2 (Speech Processing) - (COMXVIII - No.R.10) Study Group XII - Contribution No.104 Study Group XVIII - CCITT September 1982.

63. BOGNER, R.E.; FLANAGAN, J.L. "Frequency Multiplication of Speech Signals" I.E.E.E. Trans on Audio and Electro Acoustics Vol AU-17, No.3 pp202-208 September 1969.

64. CHERRY, E.C.; PHILLIPS, V.J. "Some Possible Uses of Single Sideband Signals in a Formant Tracking System" J.A.S.A. Vol - 33, No.8 ppl067-1077 August 1961.

65. GABOR, D. "Theory of Communication" Journal Institute of Electrical Engineers Vol - 93 Part III pp429-457 November 1946.

66. VILLE, J. "Theorie et Application de la Notion de Signal Analytique" Cables Et Transmission Vol - 2 pp61-74 January 1948.

67. HAYKIN, S. "Communication Systems" J. Wiley and Sons 1978.

68. BEDROSIAN, E. "The Analytic Signal Representation of Modulated Waveforms" Proceedings I.R.E. Vol - 50 pp2071-2076 October 1962.

69. OSWALD, J.R.V. "The Theory of Analytic Band-Limited Signals Applied to Carrier Systems" I.R.E. Trans on Circuit Theory Vol CT - 3, No.4 pp244-250 December 1956.

70. MARCOU, P.; DAGEUT, J "New Methods of Speech Transmission" Proceedings of Third Symposium on Information Theory London 1955.

71. DUNGUNDJI, J. "Envelopes and Pre-Envelopes of Real Waveforms" I.R.E. Trans on Information Theory Vol I.T. - 4, No.1 pp53-57 March 1958.

72. GUPTA, M.S. "Definition of Instantaneous Frequency and Frequency Measurability" American Journal of Physics Vol - 43, No.12 pp1087-1088 December 1975.

73. MANDEL, L. "Interpretation of Instantaneous Frequency" American Journal of Physics, Vol - 42 pp840-846 October 1974.

74. CORRINGTON, M.S. "Frequency Modulation Distortion caused by Common - and Adjacent - Channel Interference" R.C.A. Review Vol - 7 pp522-560 December 1946.

75. PANTER, P.F. "Modulation Noise and Spectral Analysis" McGraw - Hill 1965.

76. BROMAN, H. "The Instantaneous Frequency of a Gaussian Signal : The One-Dimensional Density Function" I.E.E.E. Trans Acoustics, Speech and Signal Processing Vol ASSP-29, No.1 pp108-111 February 1981.

77. ANGELSEN, B.A.J. "Instantaneous Frequency, Mean Frequency and Variance of Mean Frequency Estimates for Ultrasonic Blood Velocity Doppler Signals" I.E.E.E. Trans on Biomedical Engineering Vol BME-11 pp733-741 November 1981.

78. RICE, S.O. "Mathematical Analysis of Random Noise" Bell System Technical Journal Vol - 23 1944.

79. STROM, T. "On Amplitude Weighted Instantaneous Frequencies" I.E.E.E. Trans on Acoustics, Speech and Signal Processing. Vol ASSP-25, No.4 pp351-353 August 1977.

80. ROBERTS, J.H. "Angle Modulation" Peter Peregrinus Limited for I.E.E. 1977.

81. PETERSON, E. "Frequency Detection and Speech Formants" J.A.S.A. Vol - 23, No.6 pp668-674 November 1951.

82. BLACHMAN, N.M. "Zero-Crossing Rate for the Sum of Two sinusoids or a Signal Plus Noise" I.E.E.E. Trans on Information Theory pp671-675 November 1975.

83. LICKLIDER, J.C.R.; POLLACK, I. "Effects of Differentiation, Integration and Infinite Peak Clipping upon the Intelligibility of Speech" J.A.S.A. Vol - 20, No.1 pp42-51 January 1948.

84. VOELKER, H.B. "Toward a Unified Theory of Modulation Part I : Phase-Envelope Relationships" Proceedings I.E.E.E. Vol - 54, No.3 pp340-353 March 1966.

85. VOELKER, H.B. "Toward a Unified Theory of Modulation Part II : Zero Manipulation" Proceedings I.E.E.E. Vol - 54, No.5 pp735-755 May 1966.

86. VOELKER, H.B. "Demodulation of Single-Sideband Signals Via Envelope Detection" I.E.E.E. Trans on Communications Vol COM-14, No.1 pp-22-30 February 1966.

87. LOGAN, B.F.; SCHROEDER, M.R. "A Solution to Problem of Compatible Single-Sideband Transmission" I.R.E. Trans on Information Theory Vol IT-8, No.5 pp252-259 September 1962.

88. BARNARD, R.D. "On the Spectral Properties of Single-Sideband Angle Modulated Signals" B.S.T.J. pp2811-2838 November 1964.

89. KAHN, R.E.; THOMAS, J.B. "Some Bandwidth Properties of Simultaneous Amplitude and Angle Modulation" I.E.E.E. Trans on Information Theory Vol IT-11, No.4 pp516-520 October 1965.

90. VOELKER et al "On the Origin and Characteristics of Single-Sided Angle Modulation" I.E.E.E. Trans on Communications, Correspondence pp555-556 December 1965.

91. LOCKHART, G.B. "A Spectral Theory for Hybrid Modulation" I.E.E.E. Trans on Communications Vol COM-21, No.7 pp790-800 July 1973.

92. REQUICHA, A.A.G. "The Zeros of Entire Functions : Theory and Engineering Applications" Proceedings I.E.E.E. Vol - 68, No.3 pp308-328 March 1980.

93. BRISSON LOPES; LOCKHART, G.B. "Characteristics of Demodulated Narrowband SSB-FM Signals" I.E.E. Proceedings Part F, No.2 pp166-172 March 1983.

94. MARCOU, P.; DAGEUT, J. "Une Nouvelle Methode de Transmission de la Parole" Annales des Telecommunications Vol - 11, No.6 pp118-126 June 1956.

95. MAYER, H.F.; HOLZIER, E. "Verfahren und Einrichtung zur Elektrischen Nachrichtenubertragung" Patentschrift No 878.381 delivered 1, June 1953.

96. BERTHOMIER, C.; CORNILLEAU-WEHRLIN, N. "Application de la Notion de Signal Analytique a la Determination de l'Amplitude et de la Frequency Instantanee d'un Signal" Annales des Telecommunications Vol - 30, No.s 7/8 pp224-230 July/August 1975.

97. BERTHOMIER, C. "Sur une Methode d'Analyse de Signaux" Annales de Geophysique" Vol - 31, No.2 pp239-251 1975.

98. GABISON, A.; GENDRIN, R. "Appariel Analogique Destine a la Mesure en Temps Reel de l'Amplitude, de la Frequence et de la Phase Instantanees de Signaux Variant dans le Temps" Annales des Telecommunications Vol - 34, No.3/4 pp158-165 1979.

99. BRITISH PATENT SPECIFICATION 1,530,602 "High Dynamic Range Receiver for Frequency Modulated Signals".

100. VANCE, I.A.W. "An Integrated Circuit V.H.F. Radio Receiver" The Radio and Electronic Engineer Vol - 50, No.4 pp158-164 April 1980.

101. WEBB, J.A.; KELLY, M.W. "Delay Lines Help Generate Quadrature Voice for SSB" Electronics pp115 and 117 April 13, 1978.

102. OPPENHEIM, A.V.; SCHAFER, R.W. "Digital Signal Processing" Prentice-Hall Inc. pp337-367 1975.

103. SHAKER SABRI, M.; STEENAART, W. "Discrete Hilbert Transform Filtering" I.E.E.E. International Conference on Acoustics, Speech and Signal Processing. pp116-119 1976.

104. KELLY, M.W. "Echo Cancellation on Communication Circuits" Ph.D Thesis University of Canterbury Christchurch New Zealand 1979.

105. HARRIS, F.J. "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform" Proceedings I.E.E.E. Vol - 66, No.1 January 1978.

106. PRONVOST, W.; YENKIN, L.; ANDERSON, D.C.; LERNER, R. "The Voice Visualizer" American Annals of the Deaf Vol - 113 pp230-238 March 1968.

107. BARTON, G.W.; BARTON, S.H. "Forms of Sound as Shown on an Oscilloscope by Roulette Figures" Science Vol - 142, No.3589 pp1455-1456 December 1963.

108. CHANG, S.H.; PIHL, G.E.; WIREN, J "The Intervalgram as a Visual Representation of Speech Sounds" J.A.S.A. Vol - 23, No.6 pp675-679 November 1951.

109. SCARR, R.W.A. "Zero Crossings as a Means of Obtaining Spectral Information in Speech Analysis" I.E.E.E. Trans on Audio and Electro Acoustics Vol AU-16, No.2 pp247-255 June 1968.

110. ITO, M.R.; DONALDSON, R.W. "Zero Crossing Measurements for Analysis and Recognition of Speech Sounds" I.E.E.E. Trans on Audio and Electro Acoustics Vol AU-19, No.3 pp235-242 September 1971.

111. BAKER, J.M. "A New Time-Domain Analysis of Fricatives and Stop Consonants" I.E.E.E. Symposium on Speech Recognition pp134-141 April 1974.

112. FINK, L.M. Relations between the Spectrum and Instantaneous Frequency of a Signal" Problems of Information Transmission Vol - 2, No.4 pp11-21 1966.

113. RICE, S.O. "Mathematical Analysis of Random Noise (Continued)" B.S.T.J. Vol - 24 pp46-156 January 1945.

114. RICE, S.O. "Statistical Properties of a Sine Wave Plus Random Noise" B.S.T.J. Vol - 27 pp109-157 January 1948.

115. DECHAMBRE, M.; LAVERGNAT, J. "Statistical Properties of the Instantaneous Frequency for a Noisy Signal" Signal Processing 2 pp137-150 North Holland Publishing Company 1980.

116. GATKIN, N.G. et al "Probability Density of Phase Derivative of the sum of a Modulated Signal and Gaussian Noise" Radio Eng. Electron. Phys. (USSR) No.8 pp1223-1229 1965.

117. PIWNICKI, K. "Modulation Methods Related to Sine Wave Crossings" I.E.E.E. Trans on Communications Vol COM-31, No.4 pp503-508 April 1983.

118. LOGAN, B.F. "Signals Designed for Recovery after Clipping" (Parts I to III) B.S.T.J. Vol - 63, No.2 February 1984 and B.S.T.J. Vol - 63, No.4 March 1984.

119. LICKLIDER, J.C.R. "The Intelligibility of Amplitude - Dichotomized, Time-Quantized Speech Waves" J.A.S.A. Vol - 22, No.6 pp820-823 November 1950.